13 August 2004

Science

Vol. 305 No. 5686
Pages 901–1060 $10

TOWARD A HYDROGEN ECONOMY

AAAS

Special Issue

Science

Toward A Hydrogen Economy

Policymakers around the world are calling for major energy-consuming nations to move from reliance on fossil fuels to an energy economy mediated by hydrogen. A special section starting on page 957 assesses the prospects for such a transition and describes technological developments necessary for making it a reality. [Image: Stephen R. Wagner]

Volume 305
13 August 2004
Number 5686

Introduction

News

Related Editorial page 917

Review

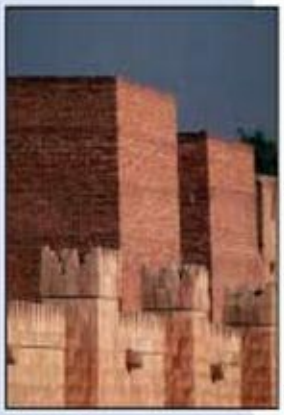
Viewpoints

Departments

932

News of the Week

News Focus



947

Letters

Books *et al.*

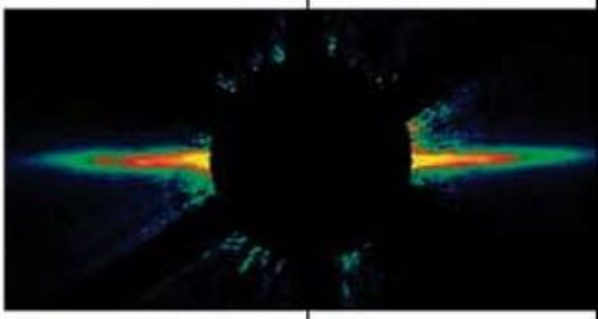
Policy Forum

Perspectives

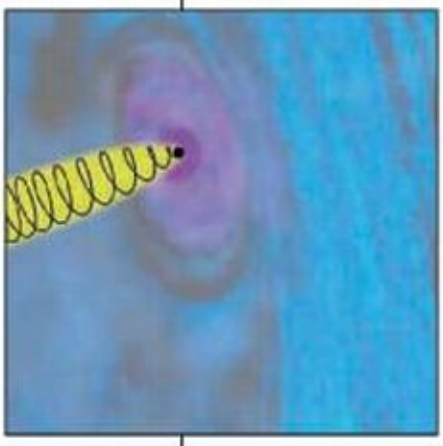
Contents continued ▶

Science

SCIENCE EXPRESS www.sciencexpress.org



BREVIA



REPORTS

931 & 978

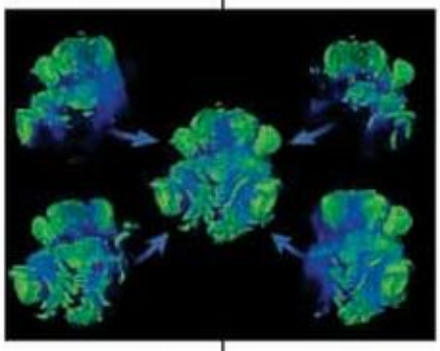
986

Contents continued

Science

REPORTS CONTINUED

997



1007

AAAS
ADVANCING SCIENCE, SERVING SOCIETY

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals Mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2004 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership and subscription (51 issues): $130 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $500. Foreign postage extra: Mexico, Caribbean (surface mail) $55; other countries (air assist delivery) $85. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. Printed in the U.S.A.

Change of address: allow 4 weeks, giving old and new addresses and 8-digit account number. Postmaster: Send change of address to Science, P.O. Box 1811, Danbury, CT 06813–1811. Single copy sales: $10.00 per issue prepaid includes surface postage; bulk rates on request. Authorization to photocopy material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act is granted by AAAS to libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that $15.00 per article is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923. The identification code for Science is 0036-8075/85 $15.00. Science is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

Contents continued ▶

sciencenow www.sciencenow.org Daily News Coverage

Resisting Ricin
New vaccine against deadly toxin shows promise in mice.

Did Climate Change Shape the Alps?
It sounds backward, but ancient erosion may explain old mountains' ruggedness.

New Light on Inflammation-Tumor Link
Key molecule on the road from colitis to colon cancer identified.

science's next wave www.nextwave.org Career Resources for Young Scientists

Sports injury rehab is more than physical.

Global/Europe: **Reach Beyond Your Dreams** *M. Arvinen-Barrow*
A specialist talks about her work with psychological rehabilitation from sports injuries.

UK: **Breaking Up with Academia** *CareerDoctor*
How should a postdoc only halfway though his contract tell his boss he's applying for nonacademic jobs?

Canada: **Canadian Science Bytes** *A. Fazekas*
Read about the latest funding, training, and job market news from Canada.

Career Development Center: **Hiring 'n' Firing—Staffing Your Lab** *J. Boss and S. Eckert*
Staffing the lab is one of the toughest challenges faced by newly independent scientists.

Career Development Center: **Navigating the Transition Award Maze** *GrantDoctor*
For experienced postdocs ready for independent careers, developing a funding strategy can be confusing.

MiSciNet: **MentorDoctor—The Clear Path** *Next Wave Staff*
The MentorDoctor team helps a minority graduate student figure out the right career choice.

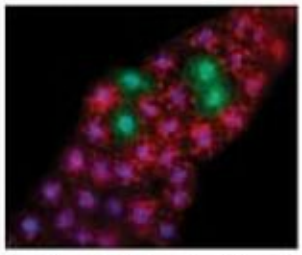
science's sage ke www.sageke.org Science of Aging Knowledge Environment

Perspective: **Ticking Fast or Ticking Slow, Through Shc Must You Go?** *F. M. Martin and J. S. Friedman*
The adapter protein $p66^{Shc}$ might integrate signals from several pathways to influence life span.

News Focus: **Fast Burn** *M. Leslie*
Genetic modification that turns fat cells into furnaces extends longevity in mice.

News Focus: **Eating Yourself to Life** *R. J. Davenport*
Signaling mechanisms that delay death induce a starvation response in fruit flies.

News Focus: **Trouncing Tau** *M. Beckman*
Mouse vaccine that neutralizes one Alzheimer's protein also eliminates another.



Turning autophagy inside out.

science's stke www.stke.org Signal Transduction Knowledge Environment

A virtual signaling community.

Forum: **E-Conference on Defining Calcium Entry Signals**
See what the experts say in this online discussion of store-operated calcium signaling.

Forum: **Open Discussion of Modeling and Computational Approaches to Cellular Signaling**
Pose a question or respond to other participants in this ongoing discussion of computational biology.

Forum: **Questions and Controversies in Zinc Signaling**
Be one of the first to discuss the role of zinc in cellular signaling processes.

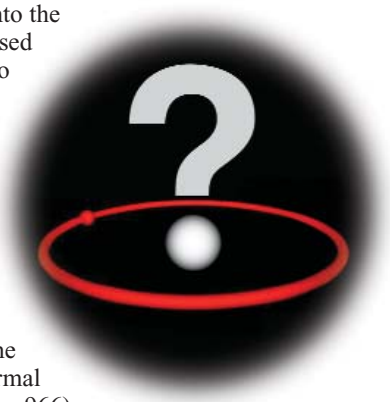
Forum: **Open Forum on Methodology**
Use this online discussion to ask questions that can help your own research.

Separate individual or institutional subscriptions to these products may be required for full-text access.

The Hydrogen Solution

If ever a phrase tripped lightly over the tongue, "the hydrogen economy" does. It appeals to the futurist in all of us, and it sounds so simple: We currently have a carbon economy that produces carbon dioxide (CO_2), the most prominent of the greenhouse gases that are warming up the world. Fortunately, however, we will eventually be able to power our cars and industries with climate-neutral hydrogen, which produces only water.

Well, can we? This issue of *Science* exposes some of the problems, and they're serious. To convert the U.S. economy in this way will require a lot of hydrogen: about 150 million tons of it in each year. That hydrogen will have to be made by extracting it from water or biomass, and that takes energy. So, at least at first, we will have to burn fossil fuels to make the hydrogen, which means that we will have to sequester the CO_2 that results lest it go into the atmosphere. That kind of dilemma is confronted in virtually all of the proposed routes for hydrogen production: We find a way of supplying the energy to create the stuff, but then we have to develop other new technologies to deal with the consequences of supplying that energy. In short, as the Viewpoint by Turner in this issue (p. 972) makes clear, getting there will be a monumental challenge.



In a recent article (*Science*, 30 July, p. 616), Secretary of Energy Spencer Abraham calls attention to the Bush administration's commitment to the hydrogen solution. The Hydrogen Fuel Initiative and FreedomCAR Partnership, announced in the 2003 State of the Union message, aims "to develop hydrogen fuel cell–powered vehicles." The United States also led the formation of the International Partnership for the Hydrogen Economy, a project in which Iceland, blessed with geothermal sources and an inventive spirit, appears to be ahead of everyone else (see p. 966). These and other initiatives are politically useful because they serve to focus public attention on the long-range goal. They rely on the premise that when the research on these new technologies is finished, we will have a better fix on the global warming problem; in the meantime, we'll put in place strictly voluntary measures to reduce CO_2 emissions. That's the case being made by the Bush administration.

The trouble with the plan to focus on research and the future, of course, is that the exploding trajectory of greenhouse gas emissions won't take time off while we are all waiting for the hydrogen economy. The world is now adding 6.5 billion metric tons of carbon to the atmosphere in the form of CO_2 annually. Some nations are cutting back on their share, but the United States, which is responsible for about a quarter of the world's total, is sticking firmly to business as usual. In each year, some of the added CO_2 will be fixed (taken up by plants in the process of photosynthesis and thus converted to biomass) or absorbed by the oceans. But because the amount added exceeds the amount removed, the concentration of atmospheric CO_2 continues to increase annually, and the added carbon remains in the atmosphere for many decades.

In fact, even if the United States and all other nations reduced the growth rate of annual emissions to zero, the concentration of greenhouse gases would continue to rise for the rest of the century, and average global temperature would increase in response. How hot it will get depends on various feedback factors: clouds, changes in Earth's reflectivity, and others. It is clear, however, that steady and significant increases in average global temperature are certain to occur, along with increases in the frequency of extreme weather events, including, as shown in the paper by Meehl and Tebaldi in this issue (p. 994), droughts and heat waves.

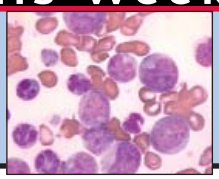
Another kind of feedback factor, of course, would be a mix of social and economic changes that might actually reduce current emissions, but current U.S. policy offers few incentives for that. Instead, it is concentrating on research programs designed to bring us a hydrogen economy that will not be carbon-free and will not be with us any time soon. Meanwhile, our attention is deflected from the hard, even painful measures that would be needed to slow our business-as-usual carbon trajectory. Postponing action on emissions reduction is like refusing medication for a developing infection: It guarantees that greater costs will have to be paid later.

Donald Kennedy
Editor-in-Chief

CREDIT: ADAPTED FROM AN ILLUSTRATION BY JEAN-FRANCOIS COLONNA (CMAP/ECOLE POLYTECHNIQUE, FT R&D)

NEWS

This Week



PAGE 929
Cancer and stem cells

930
Sue's terrible teens

MARINE EXPLORATION

NSF Takes the Plunge on a Bigger, Faster Research Sub

Deciding who will go down in history as *Alvin*'s last crew may be the biggest issue still on the table now that the U.S. government has decided to retire its famous research submarine and build a faster, roomier, and deeper diving substitute. Last week, the National Science Foundation (NSF) put an end to a decade of debate about the sub's future by announcing that it will shelve the 40-year-old *Alvin* in late 2007 and replace it with a $21.6 million craft packed with features long coveted by deep-sea scientists.

Going down. New submersible will be able to dive 6500 meters.

"It's a bittersweet moment. *Alvin* is a beloved symbol of ocean exploration," says Robert Gagosian, president of the Woods Hole Oceanographic Institution (WHOI) in Massachusetts, which operates *Alvin* and will run the new craft. "But there's a lot of excitement about the new things we'll be able to do."

The 6 August decision ended an often feisty debate over how to replace *Alvin*, which entered service in 1967 and is one of five research subs in the world that can dive below 4000 meters (*Science*, 19 July 2002, p. 326). Its storied, nearly 4000-dive career has witnessed many high-profile moments, including the discovery of sulfur-eating sea-floor ecosystems and visits to the *Titanic*. Some researchers argued for replacing the aging *Alvin* with cheaper, increasingly capable robotic vehicles. Others wanted a human-piloted craft able to reach the 11,000-meter bottom of the deepest ocean trench—far deeper than *Alvin*'s 4500-meter rating, which enables it to reach just 63% of the sea floor. Last year, after examining the issues, a National Research Council panel endorsed building a next-generation *Alvin,* but put a higher priority on constructing a $5 million robot that could dive to 7000 meters (*Science*, 14 November 2003, p. 1135).

Coming out. *Alvin*'s last dive is scheduled for late 2007.

That vehicle has yet to appear, although NSF officials say an automated sub currently under construction at WHOI partly fills the bill. And NSF and WHOI have chosen what the panel judged the riskiest approach to building a new *Alvin*: starting from scratch with a new titanium hull able to reach 6500 meters or 99% of the sea floor. The panel had suggested using leftover Russian or U.S. hulls rated to at least 4500 meters, partly because few shipyards know how to work with titanium. WHOI engineers, ▶

PATENTS

NIH Declines to March In on Pricing AIDS Drug

The National Institutes of Health (NIH) has rejected a controversial plea to use its legal muscle to rein in the spiraling cost of a widely used AIDS drug. NIH Director Elias Zerhouni last week said his agency would not "march in" and reclaim patents on a drug it helped develop because pricing issues are best "left to Congress."

The decision disappointed AIDS activists, who said it opened the door to price gouging by companies. But major research universities were quietly pleased. "This was the only decision NIH could make [based] on the law," says Andrew Neighbour, an associate vice chancellor at the University of California, Los Angeles.

The 4 August announcement was NIH's answer to a request filed in January by Essential Inventions, a Washington, D.C.–based advocacy group (*Science*, 4 June, p. 1427). It asked NIH to invoke the 1980 Bayh-Dole Act, which allows the government to reclaim patents on taxpayer-funded inventions if companies aren't making the resulting products available to the public. Specifically, the group asked NIH to march in on four patents held by Abbott Laboratories of Chicago, Illinois. All cover the anti-AIDS drug Norvir, which Abbott developed in the early 1990s with support from a 5-year, $3.5 million NIH grant.

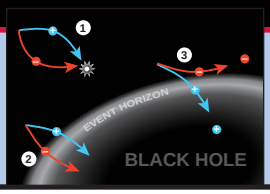
Last year, Abbott increased U.S. retail prices for some Norvir formulations by up to 400%, prompting the call for NIH to intervene and allow other manufacturers to make the drug. University groups and retired government officials who wrote the law, however, argued that such a move would be a misreading of Bayh-Dole and would undermine efforts to commercialize government-funded inventions.

In a 29 July memo, Zerhouni concluded that Abbott has made Norvir widely available to the public and "that the extraordinary remedy of march-in is not an appropriate means of controlling prices." The price-gouging charge, he added, should be investigated by the Federal Trade Commission (which is looking into the matter). Essential Inventions, meanwhile, says it will appeal to NIH's overseer, Health and Human Services Secretary Tommy Thompson. Observers doubt Thompson will intervene. –DAVID MALAKOFF

CREDITS: WOODS HOLE OCEANOGRAPHIC INSTITUTION

932 A warmer green

934 Information--please



937 A river runs through him

however, are confident that hurdle can be overcome.

Overall, the new submarine will be about the same size and shape as the current *Alvin*, so that it can operate from the existing mother ship, the *Atlantis*. But there will be major improvements.

One change is nearly 1 cubic meter more elbowroom inside the sphere that holds the pilot and two passengers. It will also offer five portholes instead of the current three, and the scientists' views will overlap with the pilot's, eliminating a long-standing complaint. A sleeker design means researchers will sink to the bottom faster and be able to stay longer. *Alvin* currently lingers about 5 hours at 2500 meters; the new craft will last up to 7 hours. A new buoyancy system will allow the sub to hover in midwater, allowing researchers to study jellyfish and other creatures that spend most of their lives suspended. And an ability to carry more weight means researchers will be able to bring more instruments—and haul more samples from the depths.

At the same time, improved electronics will allow colleagues left behind to participate in real time. As the new vehicle sinks, it will spool out a 12-kilometer-long fiber-optic cable to relay data and images. "It will put scientists, children in classrooms, and the public right in the sphere," says NSF's Emma Dieter.

Officials predict a smooth transition between the two craft. The biggest effect could be stiffer competition for time on board, because the new submersible will be able to reach areas—such as deep-sea trenches with interesting geology—once out of reach.

In the meantime, *Alvin*'s owner, the U.S. Navy (NSF will own the new craft), must decide its fate. NSF and WHOI officials will also choose a name for the new vessel, although its current moniker, taken from a 1960s cartoon chipmunk, appears to have considerable support. **–DAVID MALAKOFF**

SPACE SCIENCE

NASA Climate Satellite Wins Reprieve

Facing pressure from Congress and the White House, NASA agreed last week to rethink plans to retire a climate satellite that weather forecasters have found useful for monitoring tropical storms. The space agency said it would extend the life of the $600 million Tropical Rainfall Measuring Mission (TRMM) until the end of the year and ask the National Research Council (NRC) for advice on its future.

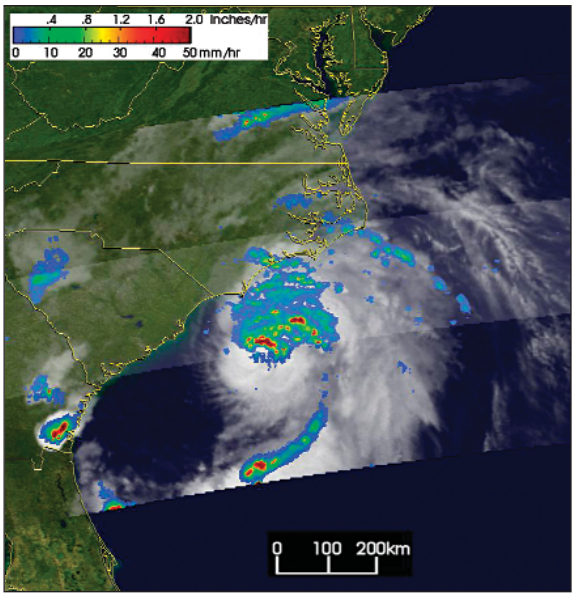
TRMM, launched on a Japanese rocket in 1997, measures rainfall and latent heating in tropical oceans and land areas that traditionally have been undersampled. Although designed for climate researchers, TRMM has also been used by meteorologists eager to improve their predictions of severe storms. "TRMM has proven helpful in complementing other satellite data," says David Johnson, director of the National Oceanic and Atmospheric Administration's (NOAA's) weather service, which relies on a fleet of NOAA spacecraft.

Climate and weather scientists protested last month's announcement by NASA that it intended to shut off TRMM on 1 August. NASA officials pleaded poverty and noted that the mission had run 4 years longer than planned. The agency said it needed to put the satellite immediately into a slow drift out of orbit before a controlled descent next spring, a maneuver that would avoid a potential crash in populated areas.

The satellite's users attracted the attention of several legislators, who complained that shutting down such a spacecraft at the start of the Atlantic hurricane season would put their constituents in danger. "Your Administration should be able to find a few tens of millions of dollars over the next 4 years to preserve a key means of improving coastal and maritime safety," chided Representative Nick Lampson (D–TX) in a 23 July letter to the White House. "A viable funding arrangement can certainly be developed between NASA and the other agencies that use TRMM's data if you desire it to happen." In an election year, that argument won the ear of the Bush Administration, in particular, NOAA Chief Conrad C. Lautenbacher Jr., who urged NASA Administrator Sean O'Keefe to rethink his decision.

On 6 August, O'Keefe said he would keep TRMM going through December. He joined with Lautenbacher in asking NRC, the operating arm of the National Academies, to hold a September workshop to determine if and how TRMM's operations should be continued. Whereas NOAA is responsible for weather forecasting, NASA conducts research and would prefer to divest itself of TRMM. "We'd be happy to give it to NOAA or a university," says one agency official. Keeping the satellite going through December will cost an additional $4 million to $5 million—"and no one has decided who is going to pay," the official added. By extending TRMM's life, NASA hopes "to aid NOAA in capturing another full season of storm data," says Ghassem Asrar, deputy associate administrator of NASA's new science directorate.

Technically, satellite operators could keep TRMM operating another 18 months, but this would come with a hidden cost. NASA would have to monitor the craft for a further 3 years before putting it on a trajectory to burn up. That option would cost about $36 million. Now that TRMM has so many highly placed friends, its supporters hope that one of them will also have deep pockets. **–ANDREW LAWLER**



Eye opener. TRMM monitored the season's first hurricane, Alex, as it approached the North Carolina coast last week.

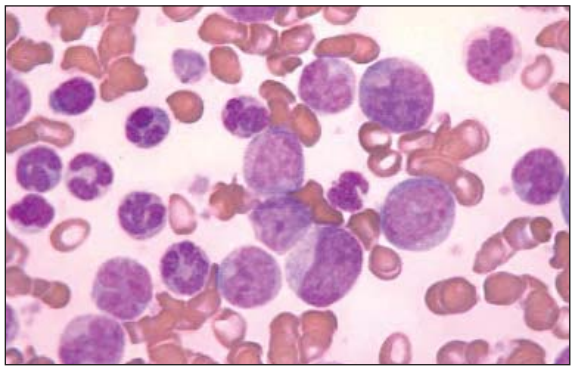
CREDIT: IMAGE PRODUCED BY HAL PIERCE (SSAI/NASA GSFC)

CANCER RESEARCH

Proposed Leukemia Stem Cell Encounters a Blast of Scrutiny

A prominent California stem cell lab says it has hit on a cadre of cells that helps explain how a form of leukemia transitions from relative indolence to life-threatening aggression. In an even more provocative claim, Irving Weissman of Stanford University and his colleagues propose in this week's *New England Journal of Medicine* that these cells, granulocyte-macrophage progenitors, metamorphose into stem cells as the cancer progresses. Some cancer experts doubt the solidity of the second claim, however.

The concept that stem cells launch and sustain a cancer has gained credence as scientists tied such cells to several blood cancers and, more recently, to breast cancer and other solid tumors (*Science*, 5 September 2003, p. 1308). Weissman's group explored a facet of this hypothesis, asking: Can nonstem cells acquire such privileged status in a cancer environment? The investigators focused on chronic myelogenous leukemia (CML), which the drug Gleevec has earned fame for treating.



Outnumbered. Immature blood cells proliferate wildly as a CML blast crisis takes hold.

The researchers gathered bone marrow samples from 59 CML patients at different stages of the disease. A hallmark of CML is its eventual shift, in patients who don't respond to early treatment, from a chronic phase to the blast crisis, in which patients suffer a massive proliferation of immature blood cells. Weissman, his colleague Catriona Jamieson, and their team noticed that among blood cells, the proportion of granulocyte-macrophage progenitors, which normally differentiate into several types of white blood cells, rose from 5% in chronic-phase patients to 40% in blast-crisis patients.

When grown in the lab, these cells appeared to self-renew—meaning that one granulocyte-macrophage progenitor spawned other functionally identical progenitor cells rather than simply giving rise to more mature daughter cells. This self-renewal, a defining feature of a stem cell, seemed dependent on the β-catenin pathway, which was previously implicated in a number of cancers, including a form of acute leukemia. Weissman and his co-authors postulate that the pathway could be a new target for CML drugs aiming to stave off or control blast crisis.

Forcing expression of β-catenin protein in granulocyte-macrophage progenitors from healthy volunteers enabled the cells to self-renew in lab dishes, the researchers report. Whereas the first stage of CML is driven by a mutant gene called *bcr-abl*, whose protein Gleevec targets, Weissman theorizes that a β-catenin surge in granulocyte-macrophage progenitors leads to the wild cell proliferation that occurs during the dangerous blast phase.

Some critics, however, say that proof can't come from the petri dish. "To ultimately define a stem cell" one needs to conduct tests in animals, says John Dick, the University of Toronto biologist who first proved the existence of a cancer stem cell in the 1990s. Studies of acute myelogenous leukemia uncovered numerous progenitor cells that seemed to self-renew, notes Dick. But when the cells were given to mice, many turned out not to be stem cells after all.

Michael Clarke of the University of Michigan, Ann Arbor, who first isolated stem cells in breast cancer, is more impressed with Weissman's results. The cells in question "clearly self-renew," he says. "The implications of this are just incredible." The suggestion that nonstem cells can acquire stemness could apply to other cancers and shed light on how they grow, he explains.

All agree that the next step is injecting mice with granulocyte-macrophage progenitors from CML patients to see whether the cells create a blast crisis. Weissman's lab is conducting those studies, and results so far look "pretty good," he says.

"What we really need to know is what cells persist in those patients" who progress to blast crisis, concludes Brian Druker, a leukemia specialist at Oregon Health & Science University in Portland. That question still tops the CML agenda, although Weissman suspects that his team has found the culprits.

–JENNIFER COUZIN

CREDIT: CDC

ScienceScope

Federal Ethics Office Faults NIH Consulting Practices

A government review of the ongoing ethics controversy at the National Institutes of Health (NIH) has found significant lapses in the agency's past procedures, according to a press report.

In a 20-page analysis, Office of Government Ethics (OGE) acting director Marilyn Glynn charges NIH with a "permissive culture on matters relating to outside compensation for more than a decade," according to excerpts in the 7 August *Los Angeles Times*. OGE reportedly found instances in which NIH lagged in approving outside consulting deals or did not approve them at all, and it concluded that some deals raised "the appearance of the use of public office for private gain." The report, addressed to the Department of Health and Human Services (HHS), also questions whether NIH officials should oversee the agency's ethics program given this spotty record. (As *Science* went to press, OGE and HHS had not released the report.)

However, the report does not recommend a blanket ban on industry consulting, according to an official who has seen it. And strict new limits proposed by NIH Director Elias Zerhouni—including no consulting by high-level employees—are consistent with the report's recommendations, says NIH spokesperson John Burklow. "We're confident that the strong policies we are developing, in addition to the steps we have already taken, will address the issues identified. We look forward to working with OGE as we finalize these policies," Burklow says.

–JOCELYN KAISER

Biopharming Fields Revealed?

The U.S. Department of Agriculture (USDA) may have to disclose the locations of biotech field trials in Hawaii after losing a round in court. The USDA issues permits for field trials of biopharmaceuticals—drug and industrial compounds produced in plants—and other genetically modified crops, but it considers the locations confidential business information. The agency is also worried about vandals.

The decision is part of a case that Earthjustice filed against USDA last year on behalf of environmental groups, arguing that field tests haven't been adequately assessed for environmental safety. Last week, a federal district court judge ruled that the field locations must be revealed to the plantiffs to assess potential harm, but gave USDA 90 days to make a stronger case against public disclosure. USDA says it is studying the decision, and Earthjustice expects the agency to appeal.

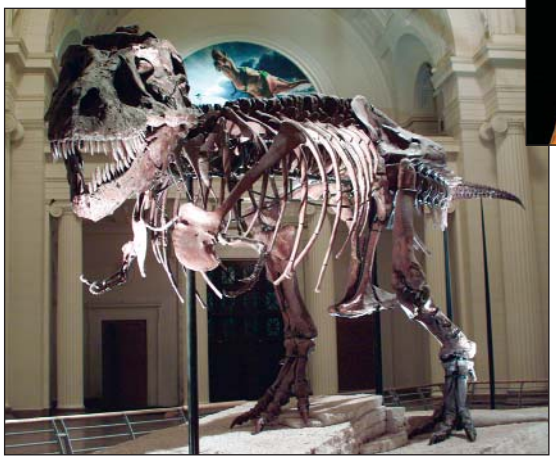
–ERIK STOKSTAD

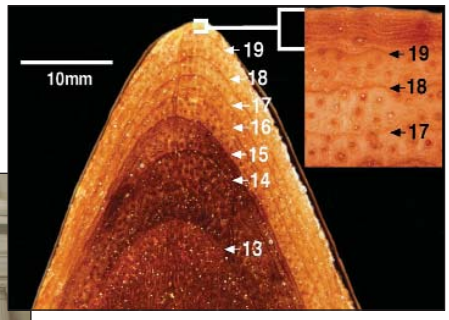
PALEONTOLOGY

Bone Study Shows *T. rex* Bulked Up With Massive Growth Spurt

Tyrannosaurus rex was a creature of superlatives. As big as a bull elephant, *T. rex* weighed 15 times as much as the largest carnivores living on land today. Now, paleontologists have for the first time charted the colossal growth spurt that carried *T. rex* beyond its tyrannosaurid relatives. "It would have been the ultimate teenager in terms of food intake," says Thomas Holtz of the University of Maryland, College Park.

Growth rates have been studied in only a half-dozen dinosaurs and no large carnivores. That's because the usual method of telling ages—counting annual growth rings in the leg bone—is a tricky task with tyrannosaurids. "I was told when I started in this field that it was impossible to age *T. rex*," recalls Gregory Erickson, a paleobiologist at Florida State University in Tallahassee, who led the study. The reason is that the weight-bearing bones of large dinosaurs become hollow with age and the internal tissue tends to get remodeled, thus erasing growth lines.





Hungry. Growth rings (*inset*) in a rib show that Sue grew fast during its teenage years.

But leg bones aren't the only place to check age. While studying a tyrannosaurid called *Daspletosaurus* at the Field Museum of Natural History (FMNH) in Chicago, Illinois, Erickson noticed growth rings on the end of a broken rib. Looking around, he found similar rings on hundreds of other bone fragments in the museum drawers, including the fibula, gastralia, and the pubis. These bones don't bear substantial loads, so they hadn't been remodeled or hollowed out.

Switching to modern alligators, crocodiles, and lizards, Erickson found that the growth rings accurately recorded the animals' ages. He and his colleagues then sampled more than 60 bones from 20 specimens of four closely related tyrannosaurids. Counting the growth rings with a microscope, the team found that the tyrannosaurids had died at ages ranging from 2 years to 28.

By plotting the age of each animal against its mass—conservatively estimated from the circumference of its femur—they constructed growth curves for each species. *Gorgosaurus* and *Albertosaurus*, both more primitive tyrannosaurids, began to put on weight more rapidly at about age 12. For 4 years or so, they added 310 to 480 grams per day. By about age 15, they were full-grown at about 1100 kilograms. The more advanced *Daspletosaurus* followed the same trend but grew faster and maxed out at roughly 1800 kilograms.

T. rex, in comparison, was almost off the chart. As the team describes this week in *Nature*, it underwent a gigantic growth spurt starting at age 14 and packed on 2 kilograms a day. By age 18.5 years, the heaviest of the lot, FMNH's famous *T. rex* named Sue, weighed more than 5600 kilograms. Jack Horner of the Museum of the Rockies in Bozeman, Montana, and Kevin Padian of the University of California, Berkeley, have found the same growth pattern in other specimens of *T. rex*. Their paper is in press at the *Proceedings of the Royal Society of London*, Series B.

It makes sense that *T. rex* would grow this way, experts say. Several lines of evidence suggest that dinosaurs had a higher metabolism and faster growth rates than living reptiles do (although not as fast as birds'). Previous work by Erickson showed that young dinosaurs stepped up the pace of growth, then tapered off into adulthood; reptiles, in contrast, grow more slowly, but they keep at it for longer. "*Tyrannosaurus rex* lived fast and died young," Erickson says. "It's the James Dean of dinosaurs."

Being able to age the animals will help shed light on the population structure of tyrannosaurids. For instance, the researchers determined the ages of more than half a dozen Albertosaurs that apparently died ▶

CREDITS: THE FIELD MUSEUM

PLANETARY SCIENCE

Los Alamos's Woes Spread to Pluto Mission

The impact of the shutdown of Los Alamos National Laboratory in New Mexico could ripple out to the distant corners of the solar system. The lab's closure last month due to security concerns (*Science*, 23 July, p. 462) has jeopardized a NASA mission to Pluto and the Kuiper belt. "I am worried," says S. Alan Stern, a planetary scientist with the Southwest Research Institute in Boulder, Colorado, who is the principal investigator.

That spacecraft, slated for a 2006 launch, is the first in a series of outer planetary flights. In those far reaches of space, solar power is not an option. Instead, the mission will be powered by plutonium-238, obtained from Russia and converted by Los Alamos scientists into pellets. But the 16 July "stand down" at the lab has shut down that effort, which already was on a tight schedule due to the lengthy review required for any spacecraft containing nuclear material.

The 2006 launch date was chosen to make use of a gravity assist from Jupiter to rocket the probe to Pluto by 2015. A 1-year delay could cost an additional 3 to 4 years in transit time. "It won't affect the science we will be able to do in a serious way, but it will delay it and introduce risks," says Stern. Some researchers fear that Pluto's thin atmosphere could freeze and collapse later in the next decade, although the likelihood and timing of that possibility are in dispute.

Los Alamos officials are upbeat. "Lab activity is coming back on line," says spokesperson Nancy Ambrosiano. Even so, last week lab director George "Pete" Nanos suspended four more employees in connection with the loss of several computer disks containing classified information, and Nanos says that it could take as long as 2 months before everyone is back at work. NASA officials declined comment, but Stern says "many people are working to find remedies."

–ANDREW LAWLER

together. They ranged in age from 2 to 20 in what might have been a pack. "You've got really young living with the really old," Erickson says. "These things probably weren't loners."

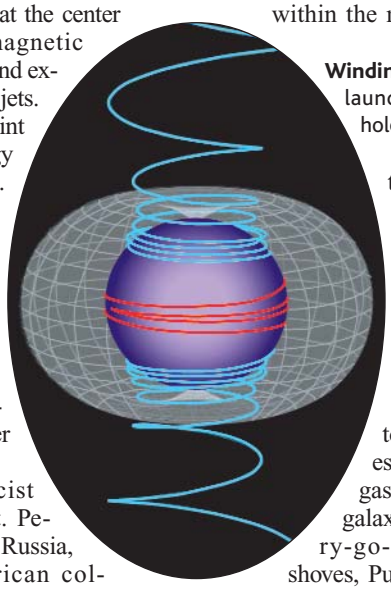
The technique could also help researchers interpret the medical history of individuals. Sue, in particular, is riddled with pathologies, and the growth rings might reveal at what age various kinds of injuries occurred. "We could see if they had a really rotten childhood or lousy old age," Holtz says. And because a variety of scrap bones can be analyzed for growth rings, more individuals can be examined. "Not many museums will let you cut a slice out of the femur of a mounted specimen," notes co-author Peter Makovicky of FMNH. "A great deal of the story about Sue was still locked in the drawers," Erickson adds. –ERIK STOKSTAD

ASTROPHYSICS

Do Black Hole Jets Do the Twist?

Among the dark secrets that nestle in galactic cores, one of the most vexing is how the gargantuan energy fountains called radio-loud quasars propel tight beams of particles and energy across hundreds of thousands of light-years. Astrophysicists agree that the power comes from supermassive black holes, but they differ sharply about how the machinery works. According to a new model, the answer might follow a familiar maxim: One good turn deserves another.

On page 978, three astrophysicists propose that a whirling black hole at the center of a galaxy can whip magnetic fields into a coiled frenzy and expel them along two narrow jets. The team's simulations paint dramatic pictures of energy spiraling sharply into space. "It has a novelty to it—it's very educational and illustrative," says astrophysicist Maurice van Putten of the Massachusetts Institute of Technology in Cambridge. But the model's simplified astrophysical assumptions allow other explanations, he says.



CREDIT: V. SEMENOV *ET AL.*

Winding up. Coiled magnetic fields launch jets from massive black holes, a model claims.

The paper, by physicist Vladimir Semenov of St. Petersburg State University, Russia, and Russian and American colleagues, is the latest word in an impassioned debate about where quasars get their spark. Some astrophysicists think the energy comes from a small volume of space around the black holes themselves, which are thought to spin like flywheels weighing a billion suns or more. Others suspect the jets blast off from blazingly hot "accretion disks" of gas that swirl toward the holes. Astronomical observations aren't detailed enough to settle the argument, and computer models require a complex mixture of general relativity, plasma physics, and magnetic fields. "We're still a few years away from realistic time-dependent simulations," says astrophysicist Ken-Ichi Nishikawa of the National Space Science and Technology Center in Huntsville, Alabama.

Semenov and his colleagues depict the churning matter near a black hole as individual strands of charged gas, laced by strong magnetic lines of force. Einstein's equations of relativity dictate the outcome, says co-author Brian Punsly of Boeing Space and Intelligence Systems in Torrance, California. The strands get sucked into the steep vortex of spacetime and tugged around the equator just outside the rapidly spinning hole, a relativistic effect called frame dragging. Tension within the magnetized ribbons keeps them intact. Repeated windings at close to the speed of light torque the stresses so high that the magnetic fields spring outward in opposite directions along the poles, expelling matter as they go.

The violent spin needed to drive such outbursts arises as a black hole consumes gas at the center of an active galaxy, winding up like a merry-go-round getting constant shoves, Punsly says. In that environment, he notes, "Frame dragging dominates everything."

Van Putten agrees, although his research suggests that parts of the black hole close to the axis of rotation also play a key role in forming jets by means of frame dragging.

Still, the basic picture—a fierce corkscrew of magnetized plasma unleashed by a frantically spinning black hole—is valuable for quasar researchers, says astrophysicist Ramesh Narayan of the Harvard-Smithsonian Center for Astrophysics in Cambridge. "This gives me a physical sense for how the black hole might dominate over the [accretion] disk in terms of jet production," he says. –ROBERT IRION

ScienceScope

Hubble Space Telescope Loses Major Instrument

One of the four main instruments on the aging Hubble Space Telescope has failed, due to an electrical fault in its power system. It will take several weeks to determine whether the Space Telescope Imaging Spectrograph (STIS) is truly deceased, but officials have slim hopes of recovery, noting that even a shuttle repair mission couldn't revive it. "It doesn't look good," says Bruce Margon, the associate director for science at the Space Telescope Science Institute in Baltimore, Maryland.

STIS, which splits incoming light into its component colors, is particularly useful for studying galaxy dynamics, diffuse gas, and black holes. Although STIS measurements account for nearly one-third of this year's Hubble science portfolio, Margon says that the telescope still has plenty of work it can do. "It will be no effort at all to keep Hubble busy," says Margon, although it is a "sad and annoying loss of capability. ... It's a bit like being a gourmet chef and being told you can never cook a chicken again."

–CHARLES SEIFE

Britain to Consider Repatriating Human Remains

The British government is requesting public comment on a proposal that could require museums and academic collections to return human remains collected around the world. Department for Culture officials last month released a white paper (www.culture.gov.uk/global/consultations) recommending that scientists identify how bones or tissues became part of their collections and seek permission from living descendants to keep identifiable remains for study. It also calls for licensing institutions that collect human remains.

Indigenous groups have long campaigned for such measures, saying that anthropologists and others have collected remains without permission. But some scientists worry that the move could harm research by putting materials out of reach and lead to expensive legal wrangles over ownership. Society needs to "balance likely harm against likely benefit," says Sebastian Payne, chief scientist at English Heritage in London, adding that "older human remains without a clear and close family or cultural relationship" are probably best left in collections. Comments are due by 29 October.

–XAVIER BOSCH

Climate researchers are finally homing in on just how bad greenhouse warming could get—and it seems increasingly unlikely that we will escape with a mild warming

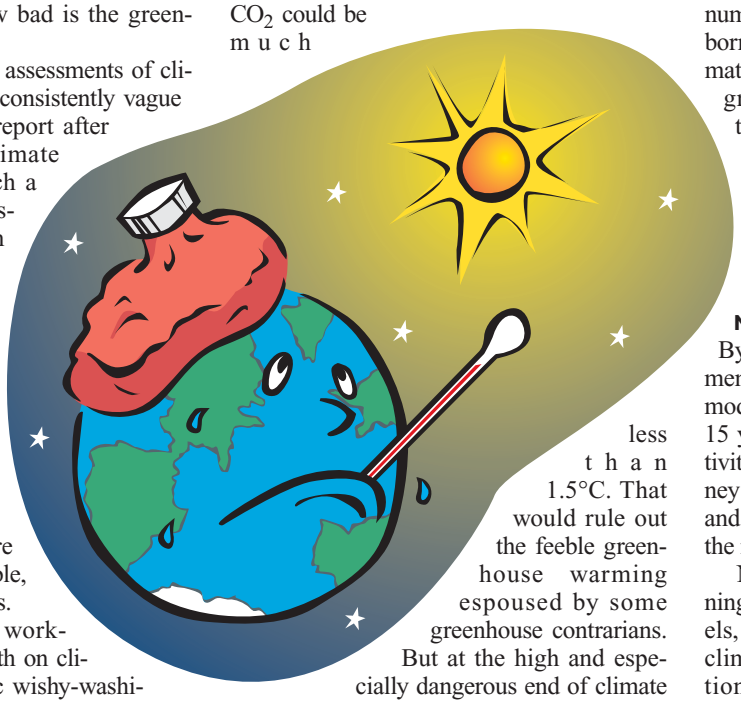
Three Degrees of Consensus

PARIS—Decades of climate studies have made some progress. Researchers have convinced themselves that the world has indeed warmed by 0.6°C during the past century. And they have concluded that human activities—mostly burning fossil fuels to produce the greenhouse gas carbon dioxide (CO_2)—have caused most of that warming. But how warm could it get? How bad is the greenhouse threat anyway?

For 25 years, official assessments of climate science have been consistently vague on future warming. In report after report, estimates of climate sensitivity, or how much a given increase in atmospheric CO_2 will warm the world, fall into the same subjective range. At the low end, doubling CO_2—the traditional benchmark—might eventually warm the world by a modest 1.5°C, or even less. At the other extreme, temperatures might soar by a scorching 4.5°, or more warming might be possible, given all the uncertainties.

At an international workshop* here late last month on climate sensitivity, climatic wishy-washiness seemed to be on the wane. "We've gone from hand waving to real understanding," said climate researcher Alan Robock of Rutgers University in New Brunswick, New Jersey. Increasingly sophisticated climate models seem to be converging on a most probable sensitivity. By running a model dozens of times under varying conditions, scientists are beginning to pin down statistically the true uncertainty of the models' climate sensitivity. And studies of natural climate changes from the last century to the last ice age are also yielding climate sensitivities.

Although the next international assessment is not due out until 2007, workshop participants are already reaching a growing consensus for a moderately strong climate sensitivity. "Almost all the evidence points to 3°C" as the most likely amount of warming for a doubling of CO_2, said Robock. That kind of sensitivity could make for a dangerous warming by century's end, when CO_2 may have doubled. At the same time, most attendees doubted that climate's sensitivity to doubled CO_2 could be much less than 1.5°C. That would rule out the feeble greenhouse warming espoused by some greenhouse contrarians. But at the high and especially dangerous end of climate sensitivity, confidence faltered; an upper limit to possible climate sensitivity remains highly uncertain.



Hand-waving climate models

As climate modeler Syukuro Manabe of Princeton University tells it, formal assessment of climate sensitivity got off to a shaky start. In the summer of 1979, the late Jule Charney convened a committee of fellow meteorological luminaries on Cape Cod to prepare a report for the National Academy of Sciences on the possible effects of increased amounts of atmospheric CO_2 on climate. None of the committee members actually did greenhouse modeling themselves, so Charney called in the only two American researchers modeling greenhouse warming, Manabe and James Hansen of NASA's Goddard Institute for Climate Studies (GISS) in New York City.

On the first day of deliberations, Manabe told the committee that his model warmed 2°C when CO_2 was doubled. The next day Hansen said his model had recently gotten 4°C for a doubling. According to Manabe, Charney chose 0.5°C as a not-unreasonable margin of error, subtracted it from Manabe's number, and added it to Hansen's. Thus was born the 1.5°C-to-4.5°C range of likely climate sensitivity that has appeared in every greenhouse assessment since, including the three by the Intergovernmental Panel on Climate Change (IPCC). More than one researcher at the workshop called Charney's now-enshrined range and its attached best estimate of 3°C so much hand waving.

Model convergence, finally?

By the time of the IPCC's second assessment report in 1995, the number of climate models available had increased to 13. After 15 years of model development, their sensitivities still spread pretty much across Charney's 1.5ºC-to-4.5ºC range. By IPCC's third and most recent assessment report in 2001, the model-defined range still hadn't budged.

Now model sensitivities may be beginning to converge. "The range of these models, at least, appears to be narrowed," said climate modeler Gerald Meehl of the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, after polling eight of the 14 models expected to be included in the IPCC' s next assessment. The sensitivities of the 14 models in the previous assessment ranged from 2.0ºC to 5.1ºC, but the span of the eight currently available models is only 2.6ºC to 4.0ºC, Meehl found.

If this limited sampling really has detected a narrowing range, modelers believe there's a good reason for it: More-powerful computers and a better understanding of atmospheric processes are making their models more realistic. For example, researchers at the Geophysical Fluid Dynamics Laboratory (GFDL) in Princeton, New Jersey, recently adopted a better way of calculating the thickness of the bottommost atmospheric layer—the boundary layer—where clouds form that are crucial to the planet's heat bal-

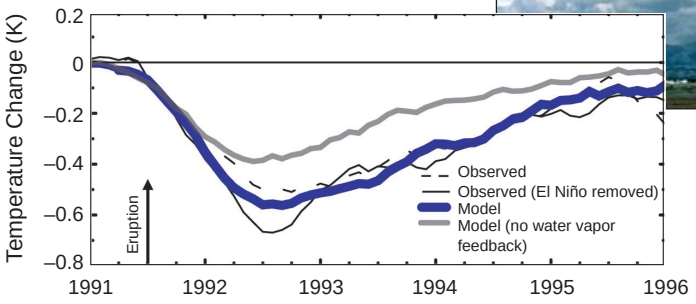
ILLUSTRATION: TIM SMITH

* Workshop on Climate Sensitivity of the Intergovernmental Panel on Climate Change Working Group I, 26–29 July 2004, Paris.

ance. When they made the change, the model's sensitivity dropped from a hefty 4.5°C to a more mainstream 2.8°C, said Ronald Stouffer, who works at GFDL. Now the three leading U.S. climate models—NCAR's, GFDL's, and GISS's—have converged on a sensitivity of 2.5°C to 3.0°C. They once differed by a factor of 2.

Less-uncertain modeling

If computer models are increasingly brewing up similar numbers, however, they sometimes disagree sharply about the physical processes that produce them. "Are we getting [similar sensitivities] for the same reason? The answer is clearly no," Jeffrey Kiehl of NCAR said of the NCAR and GFDL models. The problems come from processes called feedbacks, which can amplify or dampen the warming effect of greenhouse gases.



The biggest uncertainties have to do with clouds. The NCAR and GFDL models might agree about clouds' net effect on the planet's energy budget as CO_2 doubles, Kiehl noted. But they get their similar numbers by assuming different mixes of cloud properties. As CO_2 levels increase, clouds in both models reflect more shorter-wavelength radiation, but the GFDL model's increase is three times that of the NCAR model. The NCAR model increases the amount of low-level clouds, whereas the GFDL model decreases it. And much of the United States gets wetter in the NCAR model when it gets drier in the GFDL model.

In some cases, such widely varying assumptions about what is going on may have huge effects on models' estimates of sensitivity; in others, none at all. To find out, researchers are borrowing a technique weather forecasters use to quantify uncertainties in their models. At the workshop and in this week's issue of *Nature*, James Murphy of the Hadley Center for Climate Prediction and Research in Exeter, U.K., and colleagues described how they altered a total of 29 key model parameters one at a time—variables that control key physical properties of the model, such as the behavior of clouds, the boundary layer, atmospheric convection, and winds. Murphy and his team let each parameter in the Hadley Center model vary over a range of values deemed reasonable by a team of experts. Then the modelers ran simulations of present-day and doubled-CO_2 climates using each altered version of the model.

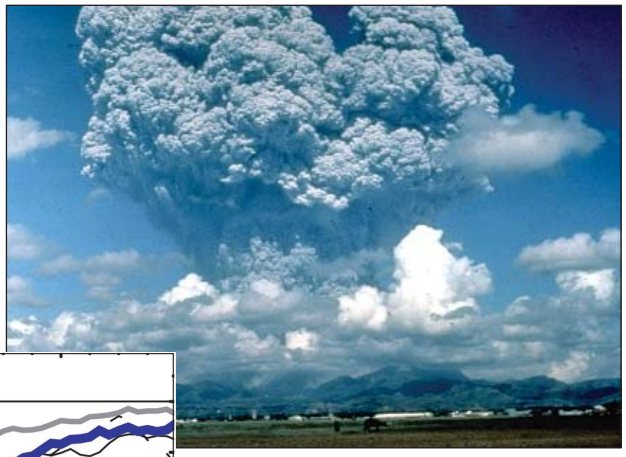
Using this "perturbed physics" approach to generate a curve of the probability of a whole range of climate sensitivities (see figure), the Hadley group found a sensitivity a bit higher than they would have gotten by simply polling the eight independently built models. Their estimates ranged from 2.4°C to 5.4°C (with 5% to 95% confidence intervals), with a most probable climate sensitivity of 3.2°C. In a nearly completed extension of the method, many model parameters are being varied at once, Murphy reported at the workshop. That is dropping the range and the most probable value slightly, making them similar to the eight-model value as well as Charney's best guess.

Murphy isn't claiming they have a panacea. "We don't want to give a sense of excessive precision," he says. The perturbed physics approach doesn't account for many uncertainties. For example, decisions such as the amount of geographic detail to build into the model introduce a plethora of uncertainties, as does the model's ocean. Like all model oceans used to estimate climate sensitivity, it has been simplified to the point of having no currents in order to make the extensive simulations computationally tractable.

Looking back

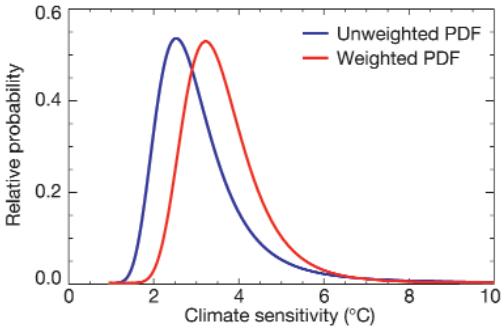
Faced with so many caveats, workshop attendees turned their attention to what may be the ultimate reality check for climate models: the past of Earth itself. Although no previous change in Earth's climate is a perfect analog for the coming greenhouse warming, researchers say modeling paleoclimate can offer valuable clues to sensitivity. After all, all the relevant processes were at work in the past, right down to the formation of the smallest cloud droplet.

One telling example from the recent past was the cataclysmic eruption of Mount Pinatubo in the Philippines in 1991. The debris it blew into the stratosphere, which stayed there for more than 2 years, was closely monitored from orbit and the ground, as was the global cooling that resulted from the debris blocking the sun. Conveniently, models show that Earth's climate system generally does not distinguish between a shift in its energy budget brought on by changing amounts of greenhouse gases and one caused by a change in the amount of solar energy allowed to enter. From the magnitude and duration of the Pinatubo cooling, climate researcher Thomas Wigley of NCAR and his colleagues have recently estimated Earth's sensitivity to a CO_2 doubling as 3.0°C. A similar calculation for the eruption of Agung in 1963 yielded a sensitivity of 2.8°C. And estimates from the five largest eruptions of the 20th century would rule out a climate sensitivity of less than 1.5°C.



Volcanic chill. Debris from Pinatubo (*above*) blocked the sun and chilled the world (*left*), thanks in part to the amplifying effect of water vapor.

Estimates from such a brief shock to the climate system would not include more sluggish climate system feedbacks, such as the expansion of ice cover that reflects radiation, thereby cooling the climate. But the globally dominant feedbacks from water vapor and clouds would have had time to work. Water vapor is a powerful greenhouse gas that's more abundant at higher temperatures, whereas clouds can cool or warm by intercepting radiant energy.



Probably warm. Running a climate model over the full range of parameter uncertainty suggests that climate sensitivity is most likely a moderately high 3.2°C (red peak).

CREDITS: (TOP TO BOTTOM) TIME LIFE PHOTOS/GETTY IMAGES; ADAPTED FROM B. SODEN *ET AL.*, *SCIENCE* **296**, 727 (2002); J. MURPHY *ET AL.*, *NATURE* **430** (2004)

More climate feedbacks come into play over centuries rather than years of climate change. So climate researchers Gabriele Hegerl and Thomas Crowley of Duke University in Durham, North Carolina, considered the climate effects from 1270 to 1850 produced by three climate drivers: changes in solar brightness, calculated from sunspot numbers; changing amounts of greenhouse gases, recorded in ice cores; and volcanic shading, also recorded in ice cores. They put these varying climate drivers in a simple model whose climate sensitivity could be varied over a wide range. They then compared the simulated temperatures over the period with temperatures recorded in tree rings and other proxy climate records around the Northern Hemisphere.

The closest matches to observed temperatures came with sensitivities of 1.5ºC to 3.0ºC, although a range of 1.0ºC to 5.5ºC was possible. Other estimates of climate sensitivity on a time scale of centuries to millennia have generally fallen in the 2ºC-to-4ºC range, Hegerl noted, although all would benefit from better estimates of past climate drivers.

The biggest change in atmospheric CO_2 in recent times came in the depths of the last ice age, 20,000 years ago, which should provide the best chance to pick the greenhouse signal out of climatic noise. So Thomas Schneider von Deimling and colleagues at the Potsdam Institute for Climate Impact Research (PIK) in Germany have estimated climate sensitivity by modeling the temperature at the time using the perturbed-physics approach. As Stefan Rahmstorf of PIK explained at the workshop, they ran their intermediate complexity model using changing CO_2 levels, as recorded in ice cores. Then they compared model-simulated temperatures with temperatures recorded in marine sediments. Their best estimate of sensitivity is 2.1ºC to 3.6ºC, with a range of 1.5ºC to 4.7ºC.

More confidence

In organizing the Paris workshop, the IPCC was not yet asking for a formal conclusion on climate sensitivity. But participants clearly believed that they could strengthen the traditional Charney range, at least at the low end and for the best estimate. At the high end of climate sensitivity, however, most participants threw up their hands. The calculation of sensitivity probabilities goes highly nonlinear at the high end, producing a small but statistically real chance of an extreme warming. This led to calls for more tests of models against real climate. They would include not just present-day climate but a variety of challenges, such as the details of El Niño events and Pinatubo's cooling.

Otherwise, the sense of the 75 or so scientists in attendance seemed to be that Charney's range is holding up amazingly well, possibly by luck. The lower bound of 1.5ºC is now a much firmer one; it is very unlikely that climate sensitivity is lower than that, most would say. Over the past decade, some contrarians have used satellite observations to argue that the warming has been minimal, suggesting a relatively insensitive climate system. Contrarians have also proposed as-yet-unidentified feedbacks, usually involving water vapor, that could counteract most of the greenhouse warming to produce a sensitivity of 0.5ºC or less. But the preferred lower bound would rule out such claims.

Most meeting-goers polled by *Science* generally agreed on a most probable sensitivity of around 3ºC, give or take a half-degree or so. With three complementary approaches—a collection of expert-designed independent models, a thoroughly varied single model, and paleoclimates over a range of time scales—all pointing to sensitivities in the same vicinity, the middle of the canonical range is looking like a good bet. Support for such a strong sensitivity ups the odds that the warming at the end of this century will be dangerous for flora, fauna, and humankind. Charney, it seems, could have said he told us so. **–RICHARD A. KERR**

Quantum Information Theory

A General Surrenders the Field, But Black Hole Battle Rages On

Stephen Hawking may have changed his mind, but questions about the fate of information continue to expose fault lines between relativity and quantum theories

Take one set of the *Encyclopedia Britannica*. Dump it into an average-sized black hole. Watch and wait. What happens? And who cares?

Physicists care, you might have thought, reading last month's breathless headlines from a conference in Dublin, Ireland. There, Stephen Hawking announced that, after proclaiming for 30 years that black holes destroy information, he had decided they don't (*Science*, 30 July, p. 586). All of which, you might well have concluded, seems a lot like debating how many angels can dance on the head of a pin.

Yet arguments about what a black hole does with information hold physicists transfixed. "The question is incredibly interesting," says Andrew Strominger, a string theorist at Harvard University. "It's one of the three or four most important puzzles in physics." That's because it gives rise to a paradox that goes to the heart of the conflict between two pillars of physics: quantum theory and general relativity. Resolve the paradox, and you might be on your way to resolving the clash between those two theories.

Eternal darkness? Spherical "event horizon" marks the region where a black hole's gravity grows so intense that even light can't escape. But is the point of no return a one-way street?

CREDIT: GSFC/NASA

Yet, as Hawking and others convince themselves that they have solved the paradox, others are less sure—and everybody is desperate to get real information about what goes on at the heart of a black hole.

The hairless hole

A black hole is a collapsed star—and a gravitational monster. Like all massive bodies, it attracts and traps other objects through its gravitational force. Earth's gravity traps us, too, but you can break free if you strap on a rocket that gets you moving beyond Earth's escape velocity of about 11 kilometers per second.

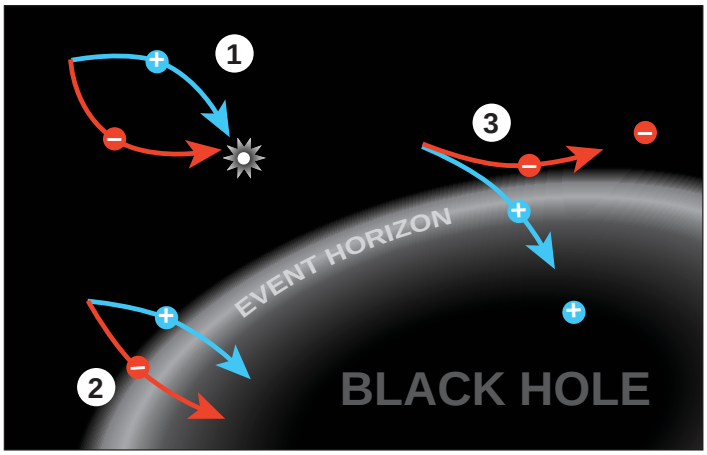
Black holes, on the other hand, are so massive and compressed into so small a space that if you stray too close, your escape velocity is faster than the speed of light. According to the theory of relativity, no object can move that fast, so nothing, not even light, can escape the black hole's trap once it strays too close. It's as if the black hole is surrounded by an invisible sphere known as an event horizon. This sphere marks the region of no return: Cross it, and you can never cross back.

The event horizon shields the star from prying eyes. Because nothing can escape from beyond the horizon, an outside observer will never be able to gather any photons or other particles that would reveal what's going on inside. All you can ever know about a black hole are the characteristics that you can spot from a distance: its mass, its charge, and how fast it's spinning. Beyond that, black holes lack distinguishing features. As Princeton physicist John Wheeler put it in the 1960s, "A black hole has no hair." The same principle applies to any matter or energy a black hole swallows. Dump in a ton of gas or a ton of books or a ton of kittens, and the end product will be exactly the same.

Not only is the information about the infalling matter gone, but information *upon* the infalling matter is as well. If you take an atom and put a message on it somehow (say, make it spin up for a "yes" or spin down for a "no"), that message is lost forever if the atom crosses a black hole's event horizon. It's as if the message were completely destroyed. So sayeth the theory of general relativity. And therein lies a problem.

The laws of quantum theory say something entirely different. The mathematics of the theory forbids information from disappearing. Particle physicists, string theorists, and quantum scientists agree that information can be transferred from place to place, that it can dissipate into the environment or be misplaced, but it can never be obliterated. Just as someone with enough energy and patience (and glue) could, in theory, repair a shattered coffee cup, a diligent observer could always reconstitute a chunk of information no matter how it's abused—even if you dump it down a black hole.

"If the standard laws of quantum mechanics are correct, for an observer outside the black hole, every little bit of information has to come back out," says Stanford University's Leonard Susskind. Quantum mechanics and general relativity are telling scientists two contradictory things. It's a paradox. And there's no obvious way out.



Cosmic refugees. Virtual particles that escape destruction near a black hole (case 3) create detectable radiation but can't carry information.

Can the black hole be *storing* the information forever rather than actually destroying it? No. In the mid-1970s, Hawking realized that black holes don't live forever; they evaporate thanks to something now known as Hawking radiation.

One of the stranger consequences of quantum theory is that the universe is seething with activity, even in the deepest vacuum. Pairs of particles are constantly winking in and out of existence (*Science*, 10 January 1997, p. 158). But the vacuum near a black hole isn't ordinary spacetime. "Vacua aren't all created equal," says Chris Adami, a physicist at the Keck Graduate Institute in Claremont, California. Near the edge of the event horizon, particles are flirting with their demise. Some pairs fall in; some pairs don't. And they collide and disappear as abruptly as they appeared. But occasionally, the pair is divided by the event horizon. One falls in and is lost; the other flies away partnerless. Without its twin, the particle doesn't wink out of existence—it becomes a real particle and flies away (see diagram). An outside observer would see these partnerless particles as a steady radiation emitted by the black hole.

Like the particles of any other radiation, the particles of Hawking radiation aren't created for free. When the black hole radiates, a bit of its mass converts to energy. According to Hawking's equations, this slight shrinkage raises the "temperature" of the black hole by a tiny fraction of a degree; it radiates more strongly than before. This makes it shrink faster, which makes it radiate more strongly, which makes it shrink faster. It gets smaller and brighter and smaller and brighter and—flash!—it disappears in a burst of radiation. This process takes zillions of years, many times longer than the present lifetime of the universe, but eventually the black hole disappears. Thus it can't store information forever.

If the black hole isn't storing information eternally, can it be letting swallowed information escape somehow? No, at least not according to general relativity. Nothing can escape from beyond the event horizon, so that idea is a nonstarter. And physicists have shown that Hawking radiation can't carry information away either. What passes the event horizon is gone, and it won't come out as the black hole evaporates.

This seeming contradiction between relativity and quantum mechanics is one of the burning unanswered questions in physics. Solving the paradox, physicists hope, will give them a much deeper understanding of the rules that govern nature—and that hold under all conditions. "We're trying to develop a new set of physical laws," says Kip Thorne of the California Institute of Technology in Pasadena.

Paradox lost

Clearly, somebody's old laws will have to yield—but whose? Relativity experts, including Stephen Hawking and Kip Thorne, long believed that quantum theory was flawed and would have to discard the no-information-destruction dictum. Quantum theorists such as Caltech's John Preskill, on the other hand, held that the relativistic view of the universe must be overlooking something that somehow salvages information from the jaws of destruction. That hope was more than wishful thinking; indeed, the quantum camp argued its case convincingly enough to sway most of the scientific community.

The clincher, many quantum and string theorists believed, lay in a mathematical correspondence rooted in a curious property of black holes. In the 1970s, Jacob Bekenstein of Hebrew University in Jerusalem and Stephen Hawking came to realize that when a black hole swallows a volume of matter, that volume can be entirely described by the

ILLUSTRATION: J. MOGILA/SCIENCE

increase of surface area of the event horizon. In other words, if the dimension of time is ignored, the essence of a three-dimensional object that falls into the black hole can be entirely described by its "shadow" on a two-dimensional object.

In the early 1990s, Susskind and the University of Utrecht's Gerard 't Hooft generalized this idea to what is now known as the "holographic principle." Just as information about a three-dimensional object can be entirely encoded in a two-dimensional hologram, the holographic principle states that objects that move about and interact in our three-dimensional world can be entirely described by the mathematics that resides on a two-dimensional surface that surrounds those objects. In a sense, our three-dimensionality is an illusion, and we are truly two-dimensional creatures—at least mathematically speaking.

Most physicists accept the holographic principle, although it hasn't been proven. "I haven't conducted any polls, but I think that a very large majority believes in it," says Bekenstein. Physicists also accept a related idea proposed in the mid-1990s by string theorist Juan Maldacena, currently at the Institute for Advanced Study in Princeton, New Jersey. Maldacena's so-called AdS/CFT correspondence shows that the mathematics of gravitational fields in a volume of space is essentially the same as the nice clean gravity-free mathematics of the boundary of that space.

Although these ideas seem very abstract, they are quite powerful. With the AdS/CFT correspondence in particular, the mathematics that holds sway upon the boundary automatically conserves information; like that of quantum theory, the boundary's mathematical framework simply doesn't allow information to be lost. The mathematical equivalence between the boundary and the volume of space means that even in a volume of space where gravity runs wild, information must be conserved. It's as if you can ignore the troubling effects of gravity altogether if you consider only the mathematics on the boundary, even when there's a black hole inside that volume. Therefore, black holes can't destroy information; paradox solved—sort of.

"String theorists felt they completely nailed it," says Susskind. "Relativity people knew something had happened; they knew that perhaps they were fighting a losing battle, but they didn't understand it on their own terms." Or, at the very least, many general relativity experts didn't think that the matter was settled—that information would still have to be lost, AdS/CFT correspondence or no. Stephen Hawking was the most prominent of the naysayers.

Paradox regained

Last month in Dublin, Hawking reversed his 30-year-old stance. Convinced by his own mathematical analysis that was unrelated to the AdS/CFT correspondence, he conceded that black holes do not, in fact, destroy information—nor can a black hole transport information into another universe as Hawking once suggested. "The information remains firmly in our universe," he said. As a result, he conceded a bet with Preskill and handed over a baseball encyclopedia (*Science*, 30 July, p. 586).

Gambling on nature. The 1997 wager among physicists Preskill, Thorne, and Hawking (*above*) became famous, but Hawking's concession (*right*) left battle lines drawn.

Despite the hoopla over the event, Hawking's concession changed few minds. Quantum and string theorists already believed that information was indestructible, thanks to the AdS/CFT correspondence. "Everybody I know in the string theory community was completely convinced," says Susskind. "What's in [Hawking's] own work is his way of coming to terms with it, but it's not likely to paint a whole new picture." Relativity experts in the audience, meanwhile, were skeptical about Hawking's mathematical method and considered the solution too unrealistic to be applied to actual, observable black holes. "It doesn't seem to me to be convincing for the evolution of a black hole where you actually *see* the black hole," says John Friedman of the University of Wisconsin, Milwaukee.

With battle lines much as they were, physicists hope some inspired theorist will break the stalemate. Susskind thinks the answer lies in a curious "complementarity" of black holes, analogous to the wave-particle duality of quantum mechanics. Just as a photon can behave like either a wave or a particle but not both, Susskind argues, you can look at information from the point of view of an observer behind the event horizon or in front of the event horizon but not both at the same time. "Paradoxes were apparent because people tried to mix the two different experiments," Susskind says.

Other scientists look elsewhere for the resolution of the paradox. Adami, for instance, sees an answer in the seething vacuum outside a black hole. When a particle falls past the event horizon, he says, it sparks the vacuum to emit a duplicate particle in a process similar to the stimulated emission that makes excited atoms emit laser light. "If a black hole swallows up a particle, it spits one out that encodes precisely the same information," says Adami. "The information is never lost." When he analyzed the process, Adami says, a key equation in quantum information theory—one that limits how much classical information quantum objects can carry—made a surprise appearance. "It simply pops out. I didn't expect it to be there," says Adami. "At that moment, I knew it was all over."

Although it might be all over for Hawking, Susskind, and Adami, it's over for different reasons—none of which has completely convinced the physics community. For the moment, at least, the black hole is as dark and mysterious as ever, despite legions of physicists trying to wring information from it. Perhaps the answer lies just beyond the horizon. **–CHARLES SEIFE**

CREDITS: (LEFT TO RIGHT) CALTECH; CHARLES SEIFE

The River Doctor

Dave Rosgen rides in rodeos, drives bulldozers, and has pioneered a widely used approach to restoring damaged rivers. But he's gotten a flood of criticism too

STOLLSTEIMER CREEK, COLORADO—"Don't be a pin-headed snarf. … Read the river!" Dave Rosgen booms as he sloshes through shin-deep water, a swaying surveying rod clutched in one hand and a toothpick in the other. Trailing in his wake are two dozen rapt students—including natural resource managers from all over the world—who have gathered on the banks of this small Rocky Mountain stream to learn, in Rosgen's words, "how to think like a river." The lesson on this searing morning: how to measure and map an abused waterway, the first step toward rescuing it from the snarfs—just one of the earthy epithets that Rosgen uses to describe anyone, from narrow-minded engineers to loggers, who has harmed rivers. "Remember," he says, tugging on the wide brim of his cowboy hat, "your job is to help the river be what it wants to be."

It's just another day at work for Rosgen, a 62-year-old former forest ranger who is arguably the world's most influential force in the burgeoning field of river restoration. Over the past few decades, the folksy jack-of-all-trades—equally at home talking hydrology, training horses, or driving a bulldozer—has pioneered an approach to "natural channel design" that is widely used by government agencies and nonprofit groups. He has personally reconstructed nearly 160 kilometers of small- and medium-sized rivers, using bulldozers, uprooted trees, and massive boulders to sculpt new channels that mimic nature's. And the 12,000-plus students he's trained have reengineered many more waterways. Rosgen is also the author of a best-selling textbook and one of the field's most widely cited technical papers—and he just recently earned a doctorate, some 40 years after graduating from college.

"Dave's indefatigable, and he's had a remarkable influence on the practice of river restoration," says Peggy Johnson, a civil engineer at Pennsylvania State University, University Park. "It's almost impossible to talk about the subject without his name coming up," adds David Montgomery, a geomorphologist at the University of Washington, Seattle.

But although many applaud Rosgen's work, he's also attracted a flood of criticism. Many academic researchers question the science underpinning his approach, saying it has led to oversimplified "cookbook" restoration projects that do as much harm as good. Rosgen-inspired projects have suffered spectacular and expensive failures, leaving behind eroded channels choked with silt and debris. "There are tremendous doubts about what's being done in Rosgen's name," says Peter Wilcock, a geomorphologist who specializes in river dynamics at Johns Hopkins University in Baltimore, Maryland. "But the people who hold the purse strings often require the use of his methods."

Class act. Dave Rosgen's system for classifying rivers is widely used in stream restoration—and detractors say commonly misused.

All sides agree that the debate is far from academic. At stake: billions of dollars that are expected to flow to tens of thousands of U.S. river restoration projects over the next few decades. Already, public and private groups have spent more than $10 billion on more than 30,000 U.S. projects, says Margaret Palmer, an ecologist at the University of Maryland, College Park, who is involved in a new effort to evaluate restoration efforts. "Before we go further, it would be nice to know what really works," she says, noting that such work can cost $100,000 a kilometer or more.

Going with the flow

Rosgen is a lifelong river rat. Raised on an Idaho ranch, he says a love of forests and fishing led him to study "all of the '-ologies' " as an undergraduate in the early 1960s. He then moved on to a job with the U.S. Forest Service as a watershed forester—working in the same Idaho mountains where he fished as a child. But things had changed. "The valleys I knew as a kid had been trashed by logging," he recalled recently. "My trout streams were filled with sand." Angry, Rosgen confronted his bosses: "But nothing I said changed anyone's mind; I didn't have the data."

Rosgen set out to change that, doggedly measuring water flows, soil types, and sediments in a bid to predict how logging and road building would affect streams. As he waded the icy waters, he began to have the first inklings of his current approach: "I realized that the response [to disturbance] varied by stream type: Some forms seemed resilient, others didn't."

In the late 1960s, Rosgen's curiosity led him to contact one of the giants of river science, Luna Leopold, a geomorphologist at the University of California, Berkeley, and a former head of the U.S. Geological Survey. Invited to visit Leopold, the young cowboy made the trek to what he still calls "Berzerkley," then in its hippie heyday. "Talk about culture shock," Rosgen says. The two men ended up poring over stream data into the wee hours.

By the early 1970s, the collaboration had put Rosgen on the path to what has become his signature accomplishment: Drawing on more than a century of research by Leopold and many others, he developed a system for lumping all rivers into a few categories based on eight fundamental characteristics, including the channel width, depth, slope, and sediment load (see graphic, p. 938). Land managers, he hoped, could use his system (there are many others) to easily classify a river and then predict how it might respond to changes, such as increased sediment. But "what started out as a

CREDIT: D. MALAKOFF/*SCIENCE*

description for management turned out to be so much more," says Rosgen.

In particular, he wondered how a "field guide to rivers" might help the nascent restoration movement. Frustrated by traditional engineering approaches to flood and erosion control—which typically called for converting biologically rich meandering rivers to barren concrete channels or dumping tons of ugly rock "rip rap" on failing banks—river advocates were searching for alternatives. Rosgen's idea: Use the classification scheme to help identify naturally occurring, and often more aesthetically pleasing, channel shapes that could produce stable rivers—that is, a waterway that could carry floods and sediment without significantly shifting its channel. Then, build it.

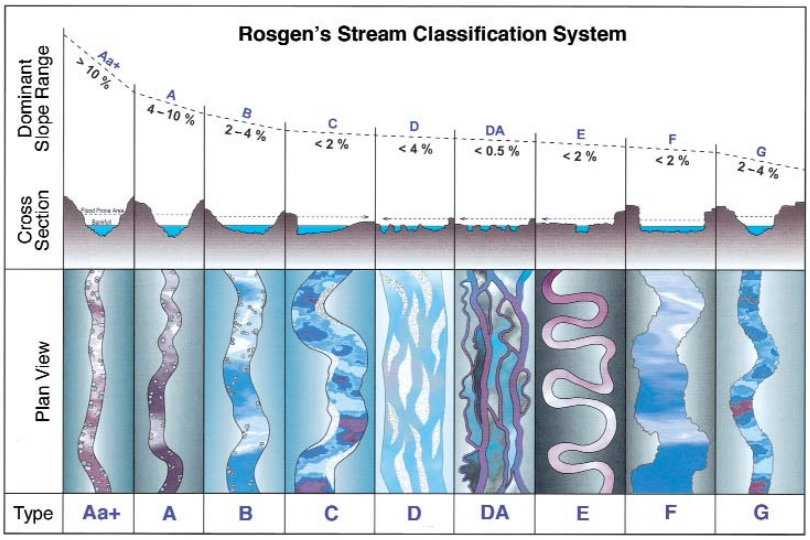
In 1985, after leaving the Forest Service in a dispute over a dam he opposed, Rosgen retreated to his Colorado ranch to train horses, refine his ideas—and put them into action. He founded a company—Wildland Hydrology—and began offering training. (Courses cost up to $2700 per person.) And he embarked on two restoration projects, on overgrazed and channelized reaches of the San Juan and Blanco rivers in southern Colorado, that became templates for what was to come.

After classifying the target reaches, Rosgen designed new "natural" channel geometries based on relatively undisturbed rivers, adding curves and boulder-strewn riffles to reduce erosion and improve fish habitat. He then carved the new beds, sometimes driving the earthmovers himself. Although many people were appalled by the idea of bulldozing a river to rescue it, the projects—funded by public and private groups—ultimately won wide acceptance, including a de facto endorsement in a 1992 National Research Council report on restoration.

Two years later, with Leopold's help, Rosgen won greater visibility by publishing his classification scheme in *Catena*, a prestigious peer-reviewed journal. Drawing on data he and others had collected from 450 rivers in the United States, Canada, and New Zealand, Rosgen divided streams into seven major types and dozens of subtypes, each denoted by a letter and a number. (Rosgen's current version has a total of 41 types.) Type "A" streams, for instance, are steep, narrow, rocky cascades; "E" channels are gentler, wider, more meandering waterways.

Although the 30-page manifesto contains numerous caveats, Rosgen's system held a powerful promise for restorationists. Using relatively straightforward field techniques—and avoiding what Rosgen calls "high puke-factor equations"—users could classify a river. Then, using an increasingly detailed four-step analysis, they could decide whether its channel was currently "stable" and forecast how it might alter its shape in response to changes, such as increased sediment from overgrazed banks. For instance, they could predict that a narrow, deep, meandering E stream with eroding banks would slowly degrade into a wide, shallow F river, then—if given enough time—restore itself back to an E. But more important, Rosgen's system held out hope of predictably speeding up the restoration process by reducing the sediment load and carving a new E channel, for instance.

The *Catena* paper—which became the basis for Rosgen's 1996 textbook, *Applied River Morphology*—distilled "decades of field observations into a practical tool," says Rosgen. At last, he had data. And people were listening—and flocking to his talks and classes. "It was an absolute revelation listening to Dave back then," recalls James Gracie of Brightwater Inc., a Maryland-based restoration firm, who met Rosgen in 1985. "He revolutionized river restoration."



A field guide to rivers. Drawing on data from more than 1000 waterways, Rosgen grouped streams into nine major types.

CREDITS: (TOP TO BOTTOM) D. MALAKOFF/*SCIENCE*; ILLUSTRATION: LEE SILVEY, FROM D. ROSGEN, *APPLIED FLUVIAL GEOMORPHOLOGY*

Rough waters

Not everyone has joined the revolution, however. Indeed, as Rosgen's reputation has grown, so have doubts about his classification system—and complaints about how it is being used in practice.

Much of the criticism comes from academic researchers. Rosgen's classification scheme provides a useful shorthand for describing river segments, many concede. But civil engineers fault Rosgen for relying on nonquantitative "geomagic," says Richard Hey, a river engineer and Rosgen business associate at the University of East Anglia in the United Kingdom. And geomorphologists and hydrologists argue that his scheme oversimplifies complex, watershed-wide processes that govern river behavior over long time scales.

Last year, in one of the most recent critiques, Kyle Juracek and Faith Fitzpatrick of the U.S. Geological Survey concluded that Rosgen's Level II analysis—a commonly used second step in his process—failed to correctly assess stream stability or channel response in a Wisconsin river that had undergone extensive study. A competing analytical method did better, they reported in the June 2003 issue of the *Journal of the American Water Resources Association*. The result suggested that restorationists using Rosgen's form-based approach would have gotten off on the wrong foot. "It's a reminder that classification has lots of limitations," says Juracek, a hydrologist in Lawrence, Kansas.

Rosgen, however, says the paper "is a pretty poor piece of work … that doesn't correctly classify the streams. … It seems like they didn't even read my book." He also emphasizes that his Level III and IV analyses are designed to answer just the kinds of questions the researchers were asking. Still, he concedes that classification may be problematic on some kinds of rivers, particularly urban waterways where massive disturbance has made it nearly impossible to make key measurements.

One particularly problematic variable, all sides agree, is "bankfull discharge," the point at which floodwaters begin to spill onto the floodplain. Such flows are believed to play a major role in determining channel form in many rivers.

Overall, Rosgen says he welcomes the critiques, although he gripes that "my most vocal critics are the ones who know the least about what I'm doing." And he recently fired back in a 9000-word essay he wrote for his doctorate, which he earned under Hey.

Rosgen's defenders, meanwhile, say the attacks are mostly sour grapes. "The academics were working in this obscure little field, fighting over three grants a year, and along came this cowboy who started getting millions of dollars for projects; there was a lot of resentment," says Gracie.

Errors on trial. Rosgen's ideas have inspired expensive failures, critics say, such as engineered meanders on California's Uvas Creek (*above*) that were soon destroyed by floods.

River revival?

The critics, however, say the real problem is that many of the people who use Rosgen's methods—and pay for them—aren't aware of its limits. "It's deceptively accessible; people come away from a week of training thinking they know more about rivers than they really do," says Matthew Kondolf, a geomorphologist at the University of California, Berkeley. Compounding the problem is that Rosgen can be a little too inspirational, adds Scott Gillilin, a restoration consultant in Bozeman, Montana. "Students come out of Dave's classes like they've been to a tent revival, their hands on the good book, proclaiming 'I believe!' "

The result, critics say, is a growing list of failed projects designed by "Rosgenauts." In several cases in California, for instance, they attempted to carve new meander bends reinforced with boulders or root wads into high-energy rivers—only to see them buried and abandoned by the next flood. In a much cited example, restorationists in 1995 bulldozed a healthy streamside forest along Deep Run in Maryland in order to install several curves—then watched the several-hundred-thousand-dollar project blow out, twice, in successive years. "It's the restoration that wrecked a river reach. … The cure was worse than the disease," says geomorphologist Sean Smith, a Johns Hopkins doctoral student who monitored the project.

Gracie, the Maryland consultant who designed the Deep Run restoration, blames the disaster on inexperience and miscalculating an important variable. "We undersized the channel," he says. But he says he learned from that mistake and hasn't had a similar failure in dozens of projects since. "This is an emerging profession; there is going to be trial and error," he says. Rosgen, meanwhile, concedes that overenthusiastic disciples have misused his ideas and notes that he's added courses to bolster training. But he says he's had only one "major" failure himself—on Wolf Creek in California—out of nearly 50 projects. "But there [are] some things I sure as hell won't do again," he adds.

What works?

Despite these black marks, critics note, a growing number of state and federal agencies are requiring Rosgen training for anyone they fund. "It's becoming a self-perpetuating machine; Dave is creating his own legion of pin-headed snarfs who are locked into a single approach," says Gillilin, who believes the requirement is stifling innovation. "An expanding market is being filled by folks with very limited experience in hydrology or geomorphology," adds J. Steven Kite, a geomorphologist at West Virginia University in Morgantown.

Kite has seen the trend firsthand: One of his graduate students was recently rejected for a restoration-related job because he lacked Rosgen training. "It seemed a bit odd that years of academic training wasn't considered on par with a few weeks of workshops," he says. The experience helped prompt Kite and other geomorphologists to draft a recent statement urging agencies to increase their training requirements and universities to get more involved (see www.geo.wvu.edu/~kite). "The bulldozers are in the water," says Kite. "We can't just sit back and criticize."

Improving training, however, is only one need, says the University of Maryland's Palmer. Another is improving the evaluation of new and existing projects. "Monitoring is woefully inadequate," she says. In a bid to improve the situation, a group led by Palmer and Emily Bernhardt of Duke University in Durham, North Carolina, has won funding from the National Science Foundation and others to undertake the first comprehensive national inventory and evaluation of restoration projects. Dubbed the National River Restoration Science Synthesis, it has already collected data on more than 35,000 projects. The next step: in-depth analysis of a handful of projects in order to make preliminary recommendations about what's working, what's not, and how success should be measured. A smaller study evaluating certain types of rock installations—including several championed by Rosgen—is also under way in North Carolina. "We're already finding a pretty horrendous failure rate," says Jerry Miller of Western Carolina University in Cullowhee, a co-author of one of the earliest critiques of Rosgen's *Catena* paper.

A National Research Council panel, meanwhile, is preparing to revisit the 1992 study that helped boost Rosgen's method. Many geomorphologists criticized that study for lacking any representatives from their field. But this time, they've been in on study talks from day one.

Whatever these studies conclude, both Rosgen's critics and supporters say his place in history is secure. "Dave's legacy is that he put river restoration squarely on the table in a very tangible and doable way," says Smith. "We wouldn't be having this discussion if he hadn't."

–DAVID MALAKOFF

CREDITS: MATTHEW KONDOLF

Letters to the Editor

Letters (~300 words) discuss material published in *Science* in the previous 6 months or issues of general interest. They can be submitted through the Web (www.submit2science.org) or by regular mail (1200 New York Ave., NW, Washington, DC 20005, USA). Letters are not acknowledged upon receipt, nor are authors generally consulted before publication. Whether published in full or in part, letters are subject to editing for clarity and space.

Virgin Rainforests and Conservation

IN REVIEWING THE HISTORY OF RAINFOREST clearance, K. J. Willis *et al.* ("How 'virgin' is virgin rainforest?", Perspectives, 16 Apr., p. 402) conclude that rainforests are "quite resilient," and that given time they "will almost certainly regenerate" from modern slash-and-burn clearance. Out of context, such statements may mislead policy-makers and weaken protection.

Although regrown rainforest may appear floristically diverse or restored (*1*), it may hold only a small proportion of the prehuman ("natural") richness and abundance of most taxa—including vertebrates, invertebrates, lichens, mosses, and microbes. Such taxa are highly dependent on the structure and microclimate of a forest (*2*, *3*). How would we know they were missing? Unfortunately, given the very poor preservation opportunities for many taxa, paleoecological evidence of the natural animal communities of rainforests is even more sparse than that for plants: The rainforests as discovered by scientists were possibly greatly impoverished compared with their prehuman state, yet we could not detect this. The prehistoric loss of the majority of the Pleistocene megafauna in some areas (e.g., giant sloths in the Amazon) means some forests can never be restored. The loss of endemic species from isolated forests is also irreversible. Few witnessing the loss of rainforest in Madagascar, for example, could believe it to be fully reversible.

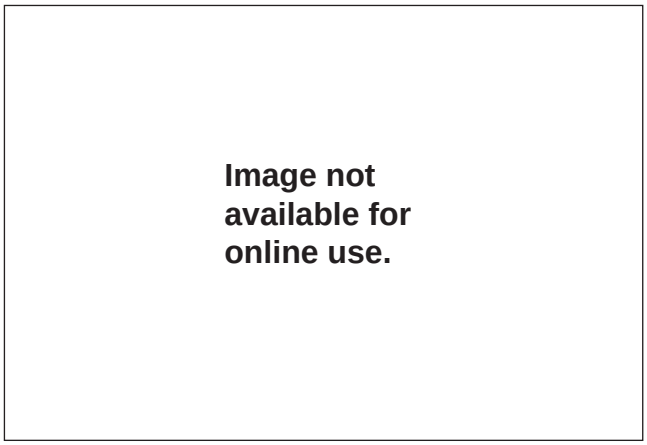
We should not assume that modern slash-and-burn clearance is comparable in impacts to that of early forest peoples—just as modern coppice management on forest reserves in Britain does not produce the same community as did "traditional" coppicing (*3*). Rainforests may be hypothesized to have been substantially impoverished by traditional management and clearance, as were British forests. Contemporary clearance—and hunting—may impoverish them further and may also be hard to monitor. A precautionary approach may be appropriate when advising forest managers.

CLIVE HAMBLER

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. E-mail: clive.hambler@zoo.ox.ac.uk

References

1. T. C. Whitmore, *An Introduction to Tropical Rain Forests* (Oxford Univ. Press, Oxford, 1998).
2. T. R. E. Southwood *et al.*, *Biol. J. Linn. Soc.* **12**, 327 (1978).
3. C. Hambler, *Conservation* (Cambridge Univ. Press, Cambridge, 2004).



Rainforest near Tari, Southern Highlands, Papua New Guinea.

CREDIT: WOLFGANG KAEHLER/CORBIS

IN THEIR PERSPECTIVE "HOW 'VIRGIN' IS virgin rainforest?" (16 Apr., p. 402), K. J. Willis *et al.* conclude that tropical humid forest regenerated quickly after the fall of prehistoric tropical societies, and that much of the "virgin" rainforest we see today is human-impacted and largely secondary. We must note that most practicing conservationists do not subscribe to the concept of "virgin" rainforest (*1*), and we disagree with the authors' suggestion that rapid rainforest regeneration may soon follow the impacts of modern development in the humid tropical forest biome (*2*).

Most prehistoric societies in the humid tropics were unlike the mechanized and industrialized societies that today dominate virtually every developing country. For example, the modern counterparts exhibit higher population densities, higher resource consumption, widespread common language, and rapid movement of the labor force in response to economic opportunities (*3*). The authors cite New Georgia in the Solomon Islands as a place where mature and species-rich "modern" forests regenerated quickly after the collapse and dispersal of large prehistoric population centers. There we find today the major impacts produced by modern industrial activities to be larger and certainly longer-lasting than the rural, traditional disturbance regimes (swidden as well as site-stable agriculture, small-scale alluvial mining, gathering of forest products, small-scale cash-cropping) that we see in modern and ancient forest societies. Today, New Georgia is beset by industrial-scale development that has seen large-scale logging lead to forest clearance for oil palm, bringing about wholesale destruction of watersheds and additional negative impacts in adjacent lagoonal coral reef ecosystems. There is little likelihood that these high-impact development zones will revert to native forest (*4*).

In Papua New Guinea, also cited by the authors, the rural customary communities inhabiting the Lakekamu Basin continually disturb the native forest through swidden agriculture, collection of a wide range of forest products, and artisanal gold-mining. However, that interior forest basin today exhibits a predominance of "mature" native rainforest, only intermittently broken by small human settlements and gardens (*5*). As with typical rural prehistoric societies, the rural subsistence human demographics of the Lakekamu produce a swidden gardening cycle that leads to rapid reforestation and minimal loss of biodiversity. Contrast this with the massive-scale development of oil palm in the fertile volcanic rainforest plains of Popondetta, about 100 km southeast of Lakekamu. There one finds large-scale monoculture that, because of its employment demands, has encouraged in-migration and a demographic shift that will, for the foreseeable future, spell intense pressure on any remaining natural forested tracts in this area. As a result, instead of regenerating humid forest, one finds continuing expansion of oil palm (as encouraged by the national government), intensive vegetable cash-cropping, and habitat degradation, which over time leads to a widespread proliferation of unproductive rank grasslands (*6*, *7*).

Overall, we see rural subsistence forest communities as forest stewards. By contrast, the large industrialized extractive industries are leading us inexorably to a world of degraded and low-biodiversity

post-forest habitats where indigenous peoples have a minimal role and no resources.

BRUCE M. BEEHLER, TODD C. STEVENSON, MICHELLE BROWN

Melanesia Center for Biodiversity Conservation, Conservation International, 1919 M Street, NW, Washington, DC 20036, USA.

References

1. J. B. Callicott, M. P. Nelson, Eds., *The Great New Wilderness Debate* (Univ. of Georgia Press, Athens, GA, 1998).
2. M. Williams, *Deforesting the Earth: From Prehistory to Global Crisis* (Univ. of Chicago Press, Chicago, IL, 2003).
3. B. Meggers, *Science* **302**, 2067 (2003).
4. E. Hviding, T. Bayliss-Smith, *Islands of Rainforest: Agroforestry, Logging and Eco-tourism in Solomon Islands* (Ashgate Press, Aldershot, UK, 2000).
5. A. Mack, Ed., *RAP Working Pap.* **9**, 1 (1998).
6. L. Curran *et al.*, *Science* **303**, 1000 (2004).
7. D. O. Fuller, T. C. Jessup, A. Salim, *Conserv. Biol.* **18**, 249 (2004).

Response

FORESTS ARE NOT MUSEUM PIECES BUT LIVING, dynamic ecosystems that have been affected by various factors—climate change, human influences, animal populations, and natural catastrophes—for millennia. The suggestion made by Hambler that tropical forests are impoverished because of prehistoric impact is not only unfounded, but also seems to imply that evidence for forest regeneration after clearance should be suppressed in case it diminishes the case for preservation. The key point that we were making is that human impact has left a lasting legacy on some areas of tropical rainforests, and the biodiverse landscapes that we value today are not necessarily pristine. In both tropical and temperate forests, there are areas in which previous human activity has enhanced biodiversity (*1*, *2*). For example, we now know that mahogany-rich forests, and the diverse flora and fauna that they support, may have originated following prehistoric catastrophic disturbance (*3*, *4*). Natural regeneration of African and Brazilian mahoganies is inhibited by the presence of more shade-tolerant rainforest tree species. In the face of increasing logging pressures, this discovery allows us to understand the steps necessary for its conservation in areas of evergreen forest—an environment in which it cannot normally regenerate (*5*).

We also argue that long-term data should be central to reexamining deforestation issues, such as that described by Hambler for Madagascar. Although there is no doubt that rapid deforestation is occurring in some areas, the process of deforestation is complex. The hypothesis that, prior to human arrival, the whole island had once been forested was overturned in the 1980s by extensive palynological work (*6*–*8*)—yet many estimates of deforestation rates in Madagascar are based on the erroneous assumption of previous 100% forest cover [e.g., (*9*)].

In response to Beehler *et al.*, we reiterate that our Perspective referred to the process of slash and burn and did not address the issue of permanent conversion of the forest following industrial-scale logging. Nor did we suggest "rapid" regeneration of forest. Indeed, the paleo-record is important in this respect because in a number of instances, it has been demonstrated that forest regeneration following clearance can take hundreds if not thousands of years.

We agree with Beehler *et al.*'s assertion that probably many conservationists working on the ground are aware that prehistoric human populations have affected currently undisturbed rainforest blocks. What they fail to mention is that this information is rarely acknowledged by the organizations for which they are working. For example, in their Web sites, major conservation organizations such as Conservation International, Wildlife Conservation Society, and the World Wildlife Fund rely on value-laden terms like "fragile," "delicate," "sensitive," and "pristine" to generate interest in rainforest projects. Although these terms certainly apply to many of the macrofauna that face extinction from commercial trade, they may be unjustified in reference to the rainforest vegetation.

The Letters of Hambler and Beehler *et al.* highlight a growing dilemma in conservation: How can long-term data on ecological resilience and variability be reconciled with a strong conservation message in the short term? We suggest that information on the long-term history of tropical rainforests can aid conservation in several ways. First, as the mahogany example highlights, management of contemporary ecosystems can be more effective if it utilizes all the ecological knowledge available. Second, providing realistic estimates of the extent and rates of forest cover change enhances the long-term credibility of the conservation movement. Such realistic estimates of the long time scales involved in the recovery of vegetation should aid those arguing for careful planning in the utilization of forest resources. Third, inevitable disturbance from rainforest exploitation should not be justification for permanent conversion of land for plantations, agriculture, cattle ranching, and mining, because long-term data highlight the potential of this biodiverse ecosystem to recover.

K. J. WILLIS, L. GILLSON, T. M. BRNCIC

Oxford Long-term Ecology Laboratory, Biodiversity Research Group, School of Geography and the Environment, Oxford, OX2 7LE UK. E-mail: kathy.willis@geog.ox.ac.uk

References

1. R. Tipping, J. Buchanan, A. Davies, E. Tisdall, *J. Biogeogr.* **26**, 33 (1999).
2. L. Kealhofer, *Asian Perspect.* **42**, 72 (2003).
3. L. J. T. White, *African Rain Forest Ecology and Conservation*, B. Weber, L. J. T. White, A. Vedder, L. Naughton-Treves, Eds. (Yale Univ. Press, New Haven, CT, 2001), p. 3.
4. L. K. Snook, *Bot. J. Linn. Soc.* **122**, 35 (1996).
5. N. D. Brown, S. Jennings, T. Clements, *Perspect. Plant Ecol. Evol. Syst.* **6**, 37 (2003).
6. D. A. Burney, *Quat. Res.* **40**, 98 (1993).
7. D. A. Burney, *Quat. Res.* **28**, 130 (1987).
8. K. Matsumoto, D. A. Burney, *Holocene* **4**, 14 (1994).
9. G. M. Green, R. W. Sussman, *Science* **248**, 212 (1990).

Stem Cell Research in Korea

LAST FEBRUARY, A GROUP OF KOREAN scientists led by W. S. Hwang and S. Y. Moon surprised the world by deriving a human embryonic stem cell line (SCNT hES-1) from a cloned blastocyst ("Evidence of a pluripotent human embryonic stem cell line derived from a cloned blastocyst," Reports, 12 Mar., p. 1669; published online 12 Feb., 10.1126/science.1094515). This is the first example of success in what might be considered a first step to human "therapeutic cloning," and it captured the attention of the world media. In response to the announcement, many have raised questions about the ethical and social environment of Korea with regard to such biotechnological investigations.

In December 2003, the Korean National Assembly passed the "Bioethics and Biosafety Act," which will go into effect in early 2005. According to the Act, human reproductive cloning and experiments such as fusion of human and animal embryos will be strictly banned [(*1*), Articles 11 and 12]. However, therapeutic cloning will be permitted in very limited cases for the cure of serious diseases. Such experiments will have to undergo review by the National Bioethics Committee (NBC) [(*1*), Article 22]. According to the Act, every researcher and research institution attempting such experiments must be registered with the responsible governmental agency [(*1*), Article 23]. Since the Act is not yet in effect, the research done by Hwang *et al.* was done without any legal control or restriction.

The Korean Bioethics Association (http://www.koreabioethics.net/), a leading bioethics group in Korea, consisting of bioethicists, philosophers, jurists, and scientists, announced "The Seoul Declaration on Human Cloning" (*2*) in 1999, demanding the ban of human reproductive cloning and the study of the socio-ethical implications of cloning research. Many nongovernment organizations and religious groups in Korea agreed with and supported the declaration.

We regret that Hwang and Moon did not wait until a social consensus about reproductive and therapeutic cloning was

achieved in Korea before performing their research. Indeed, Hwang is Chairperson of the Bioethics Committee of the Korean Society for Molecular Biology, and Moon is President of the Stem Cell Research Center of Korea and a member of its Ethics Committee. They argue that their research protocol was approved by an institutional review board (IRB). However, we are not convinced that this controversial research should be done with the approval of only one IRB. We believe that it was premature to perform this research before these issues had been resolved.

The Korean government is working to prepare regulations, guidelines, and review systems for biotechnology research in keeping with global standards (*3*). We hope that there will be no more ethically dubious research reports generated by Korean scientists before these systems are in place.

SANG-YONG SONG*

Department of Philosophy, Hanyang University, 17 Haengdang-dong, Seoul 133 -791, Korea.

*President of the Korean Bioethics Association 2002–04

References

1. Biosafety and Bioethics Act, passed 2003.
2. The Korean Bioethics Association, *J. Kor. Bioethics Assoc.* **1** (no. 1), 195 (2000).
3. Korean Association of Institutional Review Boards, Guidelines for IRB Management, 10 Feb. 2003.

Response

WE RECOGNIZE THAT OUR REPORT CHANGED the ethical, legal, and social implications of therapeutic cloning from a theoretical possibility to the first proof of principle that human embryonic stem cells can be derived from cloned blastocysts. Stem cell researchers and society at large must consider all the implications associated with therapeutic cloning. Conversations on this important topic must be all-inclusive. However, it is important to reiterate that the experiments included in our manuscript complied with all existing institutional and Korean regulations. In accordance with both Korean government regulation, as well as our own ethics, we neither have nor will conduct "human reproductive cloning and experiments such as fusion of human and animal embryos." We concur that all human embryo experiments should be overseen by appropriate medical, scientific, and bioethical experts.

In Korea, as in other countries, there is a great diversity of opinions regarding the newest scientific discoveries and when or if they should be translated into clinical research. The Korean Bioethics Association (KBA) is, in our opinion, not neutral and advocates restricting the pace of biomedical advancements, viewing new techniques as threats to society. For example, they have spoken publicly against the study of transgenic mouse models for human disease and preimplantation genetic diagnosis to help parents have healthy children. Although we respect the opinions of the KBA, we, as members of a leading Korean stem cell and cloning laboratory, are committed to discovering the medical potential of stem cells and to participating in conversations with ethical and religious groups regarding matters of bioethical concern. Our research team has always and will continue to comply with ethical regulations and any laws or guidelines promulgated by the Korean government.

WOO-SUK HWANG[1,2] AND SHIN YONG MOON[3]

[1]College of Veterinary Medicine, [2]School of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea. [3]College of Medicine, Seoul National University, Seoul, 110-744, Korea.

Changing Scientific Publishing

WE SHARE THE CONCERNS OF Y.-L. WANG *ET al.* that "[t]he direction of research is dictated more and more by publishability in high-profile journals, instead of strict

scientific considerations…" ("Biomedical Research Publication System," Letters, 26 Mar., p. 1974). We do not, however, share their conclusions, as the major components of their proposed model to improve the publication system already exist.

Wang *et al.* suggest that a post–Web publication evaluation process to determine which papers should appear in a smaller version of the printed journal that is "influenced less by haggling and more by quality" would be preferable to the current practice. In fact, this service already exists in the form of Faculty of 1000, to which we belong. The Faculty consists of over 1600 highly respected biologists, who choose and evaluate what they consider to be the best papers in their areas of biology, regardless of the journal in which the papers are published. Because this new online service evaluates each paper solely on its merits, it is beginning to make the journal in which a paper appears much less relevant.

Wang *et al.* also propose a "high-capacity Web site for posting peer-reviewed papers." This too already exists in the form of the open access site run by BioMed Central, where authors pay a flat fee to publish their research papers, which are free to be read and downloaded by anyone with access to the Web.

As these two resources are already catering to the needs delineated by Wang *et al.*, we think it makes more sense to support them, rather than to reinvent the wheel.

MARTIN C. RAFF,[1] CHARLES F. STEVENS,[2] KEITH ROBERTS,[3] CARLA J SHATZ,[4] WILLIAM T. NEWSOME[5]

[1]MRC Laboratory for Molecular Cell Biology and Cell Biology Unit, University College London, London WC1E 6BT, UK. [2]Molecular Neurobiology Laboratory, The Salk Institute of Biological Sciences, La Jolla, CA 92037, USA. [3]Department of Cell Biology, John Innes Centre, Norwich NR4 7UH, UK. [4]Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA. [5]HHMI, Department of Neurobiology, Stanford University School of Medicine, Palo Alto, CA 94305–2130, USA.

CORRECTIONS AND CLARIFICATIONS

Reports: "Three-dimensional polarimetric imaging of coronal mass ejections" by T. G. Moran and J. M. Davila (2 July, p. 66). The e-mail address for T. G. Moran on p. 67 was incorrect; the correct e-mail address is moran@orpheus.nascom.nasa.gov. Also on p. 67, a date is incorrect in the last paragraph of the second column. The correct sentence is "A halo CME was imaged three times on 29 June 1999 at 2-h intervals, and another was imaged 17 times on 4 November 1998 for 17 h at 1-h intervals." In the first complete paragraph on p. 70, the second sentence cites the wrong figure. The correct sentence is "In the topographical map (Fig. 3D), there are at least six of these linear structures visible that remain connected to the Sun, which may be legs or groups of legs of the arcade loops."

Reports: "Sites of neocortical reorganization critical for remote spatial memory" by T. Maviel *et al.* (2 July, p. 96). In the abstract, "cortex" and "cortices" were misplaced when author corrections were made to the galley. The correct sentences are as follows: "By combining functional brain imaging and region-specific neuronal inactivation in mice, we identified prefrontal and anterior cingulate cortices as critical for storage and retrieval of remote spatial memories... Long-term memory storage within some of these neocortical regions was accompanied by structural changes including synaptogenesis and laminar reorganization, concomitant with a functional disengagement of the hippocampus and posterior cingulate cortex."

Reports: "Inhibition of netrin-mediated axon attraction by a receptor protein tyrosine phosphatase" by C. Chang *et al.* (2 July, p. 103). The e-mail address given for the corresponding author, Marc Tessier-Lavigne, is incorrect. The correct e-mail address is marctl@gene.com.

ENVIRONMENT

Our Once and Future Fate

Ann Kinzig

I am penning this review—one day past due—in a plane 35,000 feet above the Atlantic. Had I followed my original plans and traveled earlier, I would have had the rare pleasure of submitting a review on time. Unfortunately, a nod to our post-9/11 world kept me out of the skies on America's Independence Day. It would somehow be comforting if we could ascribe this world to the evil or greed of a few and believe that it would be over when those few are captured or removed from office. But Paul and Anne Ehrlich's *One with Nineveh: Politics, Consumption, and the Human Future* suggests a different reality. Although not claiming to address the roots of terrorism per se, the authors make a compelling case that the combination of population growth, rampant consumption, and environmental degradation seriously threatens the livelihoods of the have-nots today and will increasingly threaten the haves in the none-too-distant future. Insecurity, hunger, and the recognition that one is entitled to a better world can breed a certain rage that will eventually find a voice.

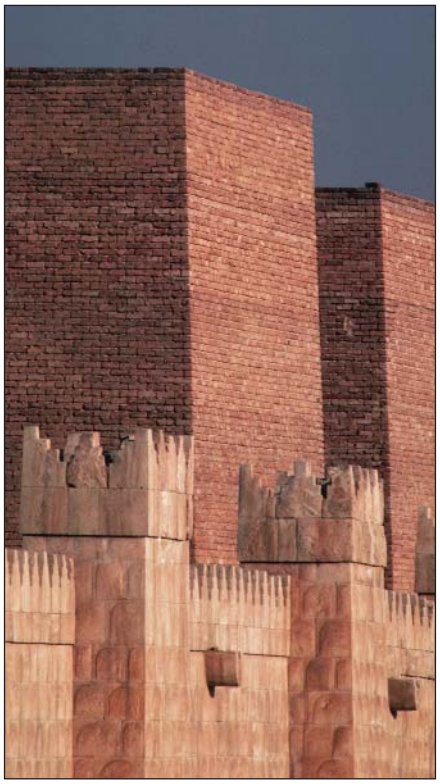
One with Nineveh
Politics, Consumption, and the Human Future
by Paul R. Ehrlich and Anne H. Ehrlich
Island Press, Washington, DC, 2004. 459 pp. $27. ISBN 1-55963-879-6.

Of course the Ehrlichs are not so naïve as to think that choreographing a better population-consumption-environment dance will rid the world of all hatred and intolerance. But surely ensuring an adequate subsistence for the poorest of the planet, and securing a sustainable future for all, would go a long way toward diminishing the power of those who preach fanaticism.

In many ways, our current environmental and human dilemma is not a new problem, as the book's title itself acknowledges. The Ehrlichs draw on a wealth of archaeological literature to document the consequences of past collisions between human aspirations and environmental limitations. We are one with Nineveh in our predilection for weakening the natural resource base that shores up the whole of human activity. However, we diverge from Nineveh in many other profound and unprecedented ways, including in our technological capacity, our global reach, and the rapidity with which we can inflict change. These differences, the Ehrlichs assert, will mean that Nineveh's fate cannot be ours. Local collapses can no longer be contained. And global rescue will require a new evolutionary step—a "conscious cultural evolution" that allows us to overcome the limitations of individual perception and formulate a more responsive societal whole.

A central thesis of the book, then, is that humanity's capacity to shape the planet has become more profound than our ability to recognize the consequences of our collective activity. The authors thoroughly document many of these consequences, such as land degradation, emerging diseases, and the loss of species. They offer some provocative insights into the causes, including limitations of the human nervous system, failures of education, and the nonlinearities in Earth systems that make effective management difficult. And they discuss potential sources for solutions: technology (which brings both promise and peril), better international institutions, and civic and religious organizations that could foment the conscious cultural evolution.

One of the joys of reading *One with Nineveh* is the sheer number of literatures the authors have reviewed. To any student of the human predicament, the bibliography alone is worth the price of the book. I particularly enjoyed the sections on economics. The Ehrlichs distill the work of many thoughtful economists to reveal some limitations of current theory, including the imperfect "rationality" of actors in the marketplace and the scaling issues that make group behavior difficult to predict from an understanding of individual preferences. More sobering, however, are the discussions of how the current theories of a few economists have driven political discourse in the wrong direction. Many contemporary economists—particularly those who have come to understand the limitations on human activity imposed by the natural environment—do not suggest that unfettered growth is a sufficient key to wealth, that markets alone can supply the necessary ingredients for a sustainable society, or that unchecked corporate activity can ensure the public good. Yet these sentiments are increasingly represented in national and international policy dialogues. More of the environmentally aware work in economics, including the collaborative work between ecologists and economists (in which the Ehrlichs regularly engage), needs to find its way into the public arena.



CREDIT: NIK WHEELER/CORBIS

Ruins at the ancient Assyrian city of Nineveh, Iraq.

Readers of *Science* should find at least two important messages in the book. The first addresses us as citizens. We are all complicit in the planet's ills, and we can all contribute to the solutions, at the very least through civic engagement and ethical reflection. The second speaks to us as scientists. There remain many unanswered questions about the functioning of our planet. As the Ehrlichs point out, science has come a long way in elucidating Earth's biogeophysical components as a complex adaptive system. Science has also advanced significantly in its understanding of the complexity of human perception and behavior across scales of social organization. We are only in the early stages of successfully joining these two perspectives to grasp how complex human dynamics engender environmental change and vice versa. There have been some steps, but more are urgently needed. Start the next leg of the journey by reading *One with Nineveh*, and see where it takes you as citizen and as scientist.

The reviewer is in the School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA. E-mail: kinzig@asu.edu

MATERIALS SCIENCE

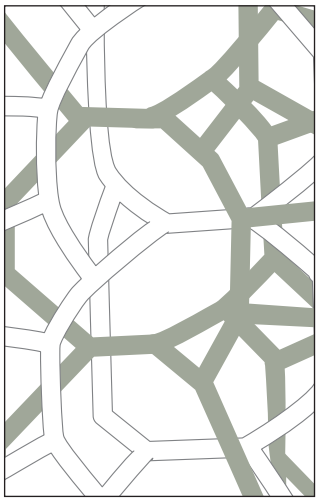
The Soft Sector in Physics

Gerard C. L. Wong

Soft matter occupies a middle ground between the solid and fluid states. These materials have neither the crystalline symmetry of solids, nor the uniform disorder of fluids. For instance, a smectic liquid crystal consists of a one-dimensional, solid-like, periodic stack of two-dimensional fluid monolayers. Liquid crystals, polymers, and colloids are commonly cited examples, but soft matter also encompasses surfactants, foams, granular matter, and networks (for example, glues, rubbers, gels, and cytoskeletons), to name a few.

The interactions that govern the behavior of soft matter are often weak and comparable in strength to thermal fluctuations. Thus these usually fragile forms of matter can respond much more strongly to stress, electric, or magnetic fields than can solid-state systems. Common themes in the behavior of soft matter include the propensity for self-organized structures (usually at length scales larger than molecular sizes), self-organized dynamics, and complex adaptive behavior (often in the form of large macroscopic changes triggered by small microscopic stimuli). These themes can be seen in a wide range of examples from the recent literature: shape-memory polymers for "smart," self-knotting surgical sutures (*1*), DNA-cationic membrane complexes in artificial gene delivery systems (*2*), colloidal crystals for templating photonic-bandgap materials (*3*), cubic lipid matrices for crystallizing integral membrane proteins (*4*), and electronic liquid crystalline phases in quantum Hall systems (*5*). (In the last case, we have come full circle, to where soft and hard condensed matter physics meet.) To a traditional condensed-matter physicist, the above list may sound at best like the animal classifications in Jorge Luis Borges's imaginary Chinese encyclopedia (*6*), but the field's broad conceptual reach is one of its strengths.

A young but already diverse field, soft condensed matter physics is expanding the province of physics in new and unexpected directions. For example, it has generated a new branch of biophysics. Most larger physics departments now have faculty who specialize in soft matter, and such materials are beginning to be covered in the undergraduate curricula in physics, chemistry, materials science, and chemical engineering. However, introducing students to the field has been a challenge because of the lack of suitable textbooks. Thus the appearance of *Structured Fluids: Polymers, Colloids, Surfactants* by Tom Witten and Phil Pincus, two pioneers in the field, is particularly welcome.



Witten and Pincus (from the physics departments at the University of Chicago and the University of California, Santa Barbara, respectively) give us a tutorial for thinking about polymers, colloids, and surfactants using a unified-scaling approach in the tradition of de Gennes's classic monograph in polymer physics (*7*). They begin with a review of statistical mechanics, and then they proceed to develop the tools needed to make simple estimates by thinking in terms of important length scales and time scales in a given phenomenon. For example: How do we estimate viscosities? How do colloids aggregate? What does a polymer look like at different length scales in different conditions, and how does that influence the way it moves? What concentrations of surfactant do we need for entangled wormlike micelles to form? Witten and Pincus demonstrate how to come up with real numbers for actual materials systems.

Structured Fluids
Polymers, Colloids, Surfactants
by Thomas A. Witten with Philip A. Pincus
Oxford University Press, Oxford, 2004. 230 pp. $74.50, £39.95. ISBN 0-19-852688-1.

Another unusual strength of the book is the authors' attention to chemical and experimental details. Too few physics textbooks explain how a polymer is made, much less mention recent synthetic strategies for controlling sequence and length with recombinant DNA technology. This book also offers an excellent, concise introduction to scattering methods, in which diffraction is presented not so much as the interference of scattered waves from atomic planes (as described in classic solid state physics textbooks) but as a Fourier transform of a density-density correlation function. This more powerful formulation facilitates generalization to diffraction from fractals and weakly ordered systems.

The authors describe a number of pedagogical "home" experiments. These cover questions including the elasticities of gels and rubber, turbidity assays, and the electrostatics of skim milk and employ such readily available household components as gelatin, rubber bands, and laser pointers. Many interesting concepts are relegated to the appendices, which reward careful reading. These range from a consideration of the dilational invariance of random walks to a presentation of the celebrated Gauss-Bonnet theorem (which seems as much a miracle as it is differential geometry).

The book's fairly short length required the authors to make hard choices. As a result, the coverage is uneven and there are notable omissions. (For example, the rotational-isomerization-state model for polymer conformations is only discussed qualitatively, as are semiflexible chains.) In addition, readers would benefit from having more worked problems. On the other hand, the book is very readable, and it can be easily adapted for a one-semester or a one-quarter course. Instead of opting for an encyclopedic treatment, Witten and Pincus cultivate a physicist's style of thought and intuition, which often renders knowledge weightless. *Structured Fluids* belongs on one's shelf beside recent works by Paul Chaikin and Tom Lubensky (*8*), Jacob Israelachvili (*9*), and Ronald Larson (*10*). These books rectify and expand prevailing notions of what condensed matter physics can be.

The reviewer is in the Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, 1304 West Green Street, Urbana, IL 61801, USA. E-mail: gclwong@uiuc.edu

References and Notes

1. A. Lendlein, R. Langer, *Science* **296**, 1673 (2002).
2. Y. A. Vlasov, X. Z. Bo, J. Z. Sturn, D. J. Norris, *Nature* **393**, 550 (1998).
3. J. O. Rädler, I. Koltover, T. Salditt, C. R. Safinya, *Science* **275**, 810 (1997).
4. E. Pebay-Peyroula, G. Rummel, J. P. Rosenbusch, E. M. Landau, *Science* **277**, 1676 (1997).
5. S. A. Kivelson, E. Fradkin, V. J. Emery, *Nature* **393**, 550 (1998).
6. Borges describes "a certain Chinese encyclopedia called the *Heavenly Emporium of Benevolent Knowledge*. In its distant pages it is written that animals are divided into (a) those that belong to the Emperor; (b) embalmed ones; (c) those that are trained; (d) suckling pigs; (e) mermaids; (f) fabulous ones; (g) stray dogs; (h) those that are included in this classification; (i) those that tremble as if they were mad; (j) innumerable ones; (k) those drawn with a very fine camel's hair brush; (l) et cetera; (m) those that have just broken the flower vase; (n) those that at a distance resemble flies." J. L. Borges, *Selected Non-Fictions*, E. Weinberger, Ed. (Penguin, New York, 1999), pp. 229–232.
7. P.-G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, Ithaca, NY, 1979).
8. P. M. Chaikin, T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge Univ. Press, Cambridge, 1995).
9. J. N. Israelachvili, Ed., *Intermolecular and Surface Forces* (Academic Press, London, ed. 2, 1992).
10. R. G. Larson, *The Structure and Rheology of Complex Fluids* (Oxford Univ. Press, Oxford, 1999).

CREDIT: ADAPTED FROM WITTEN AND PINCUS, 2004

ETHICS

Human Health Research Ethics

E. Silbergeld, S. Lerman,* L. Hushka

The issue of ethics surrounding studies for regulatory decision–making has been the subject of recent discussions at the Environmental Protection Agency (EPA) that could have broad implications for human subject research. In 2000, a report from a joint meeting of the Agency's Science Advisory Board (SAB) and the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA) Science Advisory Panel (SAP) recommended that the Agency require "active and aggressive" review of human studies conducted by external groups (*1*). EPA announced a moratorium indicating it would not consider "third-party" generated data (i.e., from academia, industry, or public interest groups) in its regulatory process until ethical issues were resolved (*2*). This ban centered on several clinical studies submitted by pesticide manufacturers since 1998. However, EPA's policy appeared to have implications for other toxicology and epidemiology studies. In 2001, EPA requested that the National Research Council (NRC) "furnish recommendations regarding the particular factors and criteria EPA should consider to determine the potential acceptability of third-party studies." EPA also asked the NRC to provide advice on a series of questions, including "recommendations on whether internationally accepted protocols for the protection of human subjects (the 'Common Rule') could be used to develop scientific and ethical criteria for EPA" (*3*).

In May 2003, EPA issued an Advanced Notice of Proposed Rulemaking (ANPRM), the first formal step toward developing a regulatory standard and solicited public comment (*4*). The ANPRM noted that third-party research is not legally subject to the Common Rule. The Common Rule, which is administered by the Department of Health and Human Services (DHHS), details accepted ethical standards for the protection of human subjects in research conducted or sponsored by all federal agencies (*5*). In its ANPRM, EPA raised questions regarding policy options being considered, including applicability of the Common Rule and whether the standard of acceptability should vary depending on research design, provenance, impact on regulatory standard, or EPA's assessment of the risks and benefits of the research. In addition, they requested input on a prospective and retroactive study review process.

We do not find a compelling reason for EPA to propose alternate and complex criteria. We believe that the best approach is the application of the Common Rule or equivalent international standards (*6*, *7*). The Common Rule codifies existing ethical guidance, is built on decades of experience and practice, and thus is both necessary and sufficient to ensure protection of human research subjects. There should be no difference in the standards based on the study design, source of funding, or, most disturbingly, the impact of the study on a regulatory standard. Otherwise, data that were obtained in studies deemed ethically acceptable under the Common Rule could be excluded, or (perhaps worse) data from studies that do not meet these norms could be included.

We find troubling the notion that the ethical standard for a human toxicity test or a clinical trial would be different when conducted by a nonprofit organization or an industry. Whether or not studies with human subjects to test pesticides and industrial chemicals will be judged ethically acceptable is not the point. We are also concerned that different ethical norms might be applied on the basis of whether the study's conclusions strengthen or relax an EPA regulatory position. Biasing the process in either direction is bad science and public policy.

In February 2004, the NRC recommended (*8*) that studies be conducted and used for regulatory purposes if they are adequately designed, societal benefits of the study outweigh any anticipated risks, and recognized ethical standards and procedures are observed. It also stated that EPA should ensure that all research it uses is reviewed by an appropriately constituted Institutional Review Board (IRB) before initiation, regardless of the source of funding. These conclusions are consistent with other counsel that all research proposals involving human subjects be submitted for scientific and ethical review (*9*).

Although we agree with these recommendations, we strongly disagree with NRC's call for creation of an EPA review process and review board for human studies proposed for use in formulating regulations. Private entities would submit research plans before beginning a study, and again before submitting the study results. It is unclear how post-study review can contribute to protection of research subjects. Introduction of such a parallel review process will create confusion regarding which set of rules applies to a particular study. It is also likely to create resource and logistical problems. We suggest that EPA require that private entities obtain review under the Common Rule or its foreign equivalent before undertaking a study and provide documentation of this review in order to submit their data for regulatory purposes. By requiring studies to follow the Common Rule or a foreign equivalent, EPA can strongly discourage the practice of conducting human-subjects research and clinical trials outside the United States, to avoid federal scrutiny.

By a strong endorsement and legally binding adoption of the Common Rule and equivalent international standards, EPA can ensure that ethical concerns are fully considered. By joining the community of biomedical ethics, rather than establishing a separate path, EPA will strengthen all of our efforts.

E. K. Silbergeld is with the Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD 21205, USA. S. E. Lerman is with ExxonMobil Biomedical Sciences Inc., Annandale, NJ 08801, USA. L. J. Hushka is with Exxon Mobil Corporation, Houston TX 77079, USA.

*Author for correspondence. E-mail: steven.e.lerman@exxonmobil.com

References and Notes

1. Science Advisory Board and the FIFRA Scientific Advisory Panel, EPA, "Comments on the use of data from the testing of human subjects" (EPA-SAB-EC-00-017, EPA, Washington, DC, 2000).
2. EPA, Agency requests National Academy of Sciences input on consideration of certain human toxicity studies; announces interim policy (press release, 14 December 2001).
3. National Research Council (NRC), *Use of Third-Party Toxicity Research with Human Research Participants* (National Academies Press, Washington, DC, 2002).
4. EPA, Human testing; Advance notice of proposed rulemaking, Docket no. OPP-2003-0132, *Fed. Regist.* **68**, 24410 (2003).
5. DHHS, Protection of human subjects, Code of Federal Regulations (CFR) **40**, part 26 (2001).
6. World Medical Association, "Declaration of Helsinki: Ethical principles for medical research involving human subjects" (World Medical Association, Edinburgh, 2000).
7. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH Topic E6: Guideline for Good Clinical Practice, Geneva, 1996).
8. NRC, *Intentional Human Dosing Studies for EPA Regulatory Purposes: Scientific and Ethical Issues* (National Academies Press, Washington, DC, 2004).
9. The Council for International Organizations of Medical Sciences (CIOMS), *International Ethical Guidelines for Biomedical Research Involving Human Subjects* (National Academies Press, Washington, DC, 2002).

PHYSICS

Half Full or Half Empty?

J. P. Eisenstein

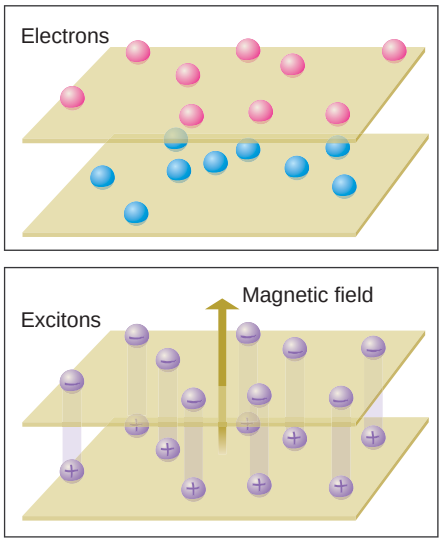
Only a few years after Bardeen, Cooper, and Schrieffer introduced their successful theory of superconductivity in metals (*1*, *2*), the idea that something similar might happen in semiconductors was advanced (*3*). Electrons in a superconductor, even though they repel one another, join to form pairs. Known as Cooper pairs, these composite objects are members of a class of quantum particles called bosons. Unlike individual electrons and the other members of the particles called fermions, bosons are not bound by the Pauli exclusion principle: Any number of bosons can condense into the same quantum state. Bose condensation is at the root of the bizarre properties of superfluid helium and is nowadays being intensely studied in ultracold atomic vapors. The condensation of Cooper pairs in a metal leads not only to the well-known property of lossless conduction of electricity, but also to a variety of other manifestations of quantum mechanics on a macroscopic scale.

In a semiconductor, there are both electrons and holes. Holes are unfilled electron states in the valence band of the material. Remarkably, holes behave in much the same way as electrons, with one crucial difference: Their electrical charge is positive rather than negative. Electrons and holes naturally attract one another, and thus pairing seems very likely. Like Cooper pairs, these excitons, as they are known, are bosons. If a suitably dense collection of excitons could be cooled to a sufficiently low temperature, Bose condensation ought to occur and a new state of matter should emerge. Or so went the thinking in the early 1960s.

Alas, there is a problem: Excitons are unstable. They typically survive only about a nanosecond before the electron simply falls into the hole, filling the empty valence band state and giving birth to a flash of light in the process. A nanosecond is not very long, and this left the prospects for creating a condensate of excitons in a bulk semiconductor pretty poor. Over the last decade the situation has improved considerably through the use of artificial semiconductor structures in which the electrons and holes are confined to thin slabs of material separated by a thin barrier layer. This physical separation slows the recombination substantially, and some very interesting, and provocative, results have been obtained (*4*–*6*). Excitonic Bose condensation has, however, remained elusive.

Last March, experimental results reported at the meeting of the American Physical Society in Montreal by independent California Institute of Technology/Bell Labs and Princeton groups have revealed clear signs of excitonic Bose condensation (*7*, *8*). Remarkably, however, the findings were made with samples consisting of two layers of electrons or two layers of holes. How can one have exciton condensation without electrons and holes in the same sample? The trick is to use a large magnetic field to level the playing field between electron-hole, electron-electron, and hole-hole double-layer systems (see the figure on this page).

Suppose that only electrons are present in a thin layer of semiconductor. (This can easily be achieved by doping with a suitable impurity.) Applying a large magnetic field perpendicular to this system creates a ladder of discrete energy levels for these electrons to reside in. If the field is large enough, the electrons may only partially fill the lowest such level. Now, borrowing the old viticultural metaphor, is the level partially filled or partially empty? The magnetic field allows us to choose either point of view. If it is the latter, we may think of the system as a collection of holes, just as we always do with a partially filled valence band in a semiconductor. Now bring in a second identical layer of electrons, and position it parallel to the first. We remain free to take either the partially full or partially empty point of view with this layer. Let us consider the first layer in terms of holes and the second in terms of electrons. If the layers are close enough together, the holes and electrons will bind to each other because of their mutual attraction to form interlayer excitons. All we need to do is ensure that there are no electrons or holes left over. A moment's thought shows that the way to do this is to ensure that the total number of electrons in both of the original layers is just enough to completely fill precisely one of the energy levels created by the magnetic field. This is easily done by adjusting the magnetic field strength to the right value (*9*–*11*).

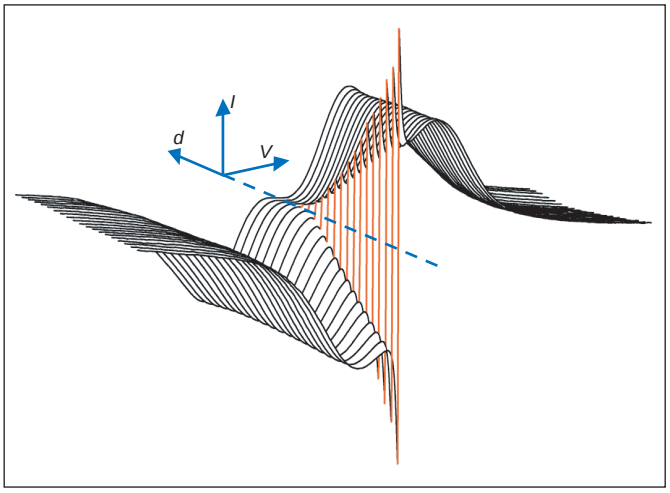


Stabilizing excitons. With the application of a strong magnetic field to a double layer (**top**) of electrons in a semiconductor, electron-hole pairs (excitons) can be stabilized against decay and undergo Bose condensation (**bottom**).

An immense advantage of electron-electron or hole-hole double-layer systems for creating exciton condensates is that they are in equilibrium. In the electron-electron case, only the conduction band of the semiconductor is involved. In the hole-hole case, it is only the valence band. No optical recombination occurs in either system. Experimenters can proceed at their leisure.

The new results reported in Montreal clearly reveal that electrons and holes are binding to each other to form electrically neutral pairs. To demonstrate this, a variation on a time-honored electrical measurement was performed. When an electrical current flows at right-angles to a magnetic field, the Lorentz force on the carriers leads to a voltage perpendicular to both the field and the current. This is the famous Hall effect. One of the most important aspects of the Hall effect is that the sign of the Hall voltage is determined by the sign of the charge of the particles carrying the current. In the recent experiments, equal but oppositely directed electrical currents were made to flow through the two layers of electrons (or holes). This was done because a uniform flow of excitons in one direction, if present, would necessarily involve oppositely directed electrical currents in the two layers. Meanwhile, the Hall voltage in one of the layers was monitored. Normally one would expect that the sign of this voltage would be

The author is at the California Institute of Technology, Pasadena, CA 91125, USA. E-mail: jpe@caltech.edu



Exciton condensation. Onset of exciton condensation as detected in the current, which quantum mechanically tunnels between the two layers in the double layer two-dimensional electron system, as a function of the interlayer voltage *V*. A family of curves is shown, each one for a different effective separation *d* between the layers. At large *d*, the tunneling current near $V = 0$ is strongly suppressed. As *d* is reduced, however, an abrupt jump in the current (highlighted in red) develops around $V = 0$. This jump, reminiscent of the Josephson effect in superconductivity, is a compelling indicator of the expected quantum coherence in the excitonic state.

determined by the sign of the charge carriers only in the layer being measured. What the California Institute of Technology/Bell Labs team and the researchers at Princeton found was that under the conditions in which exciton condensation was expected, the Hall voltage simply vanished. The explanation for this is simple: The oppositely directed currents in the two layers are being carried not by individual particles, but by interlayer excitons. Excitons have no net charge and so there is no net Lorentz force on them, and hence no Hall voltage develops.

A vanishing Hall voltage is compelling evidence that excitons are present. By itself, however, it does not prove that the exciton gas possesses the kind of long-range quantum coherence expected of a Bose condensate. Although both groups also found that the conductivity of the exciton gas appears to diverge as the temperature approaches absolute zero, an independent indicator of coherent behavior would make a much more compelling case. Interestingly, prior experiments by the California Institute of Technology/Bell Labs group provided just such an indication (*12*). These earlier experiments revealed a gigantic enhancement of the ability of electrons to quantum mechanically "tunnel" through the barrier separating the layers under the conditions in which exciton condensation was expected (see the figure on this page). Taken together, the new Hall effect measurements and the older tunneling studies very strongly suggest that the vision of excitonic Bose condensation first advanced some 40 years ago has finally been achieved.

References

1. J. Bardeen, L. N. Cooper, J. R. Schrieffer, *Phys. Rev.* **106**, 162 (1957).
2. J. Bardeen, L. N. Cooper, J. R. Schrieffer, *Phys. Rev.* **108**, 1175 (1957).
3. L. V. Keldysh, Y. V. Kopaev, *Fiz. Tverd. Tela. (Leningrad)* **6**, 2791 (1964) [*Sov. Phys.* **6**, 2219 (1965)].
4. D. B. Snoke, *Science* **298**, 1368 (2002).
5. L. V. Butov, *Solid State Commun.* **127**, 89 (2003).
6. C. W. Lai, J. Zoch, A. C. Gossard, D. S. Chemla, *Science* **303**, 503 (2004).
7. M. Kellogg, J. P. Eisenstein, L. N. Pfeiffer, K. W. West, *Phys. Rev. Lett.* **93**, 036801 (2004).
8. E. Tutuc, M. Shayegan, D. Huse, *Phys. Rev. Lett.* **93**, 036802 (2004).
9. H. Fertig, *Phys. Rev. B* **40**, 1087 (1989).
10. E. H. Rezayi, A. H. MacDonald, *Phys. Rev. B* **42**, 3224 (1990).
11. X. G. Wen, A. Zee, *Phys. Rev. Lett.* **69**, 1811 (1992).
12. I. B. Spielman, J. P. Eisenstein, L. N. Pfeiffer, K. W. West, *Phys. Rev. Lett.* **84**, 5808 (2000).

NEUROSCIENCE

Addicted Rats

Terry E. Robinson

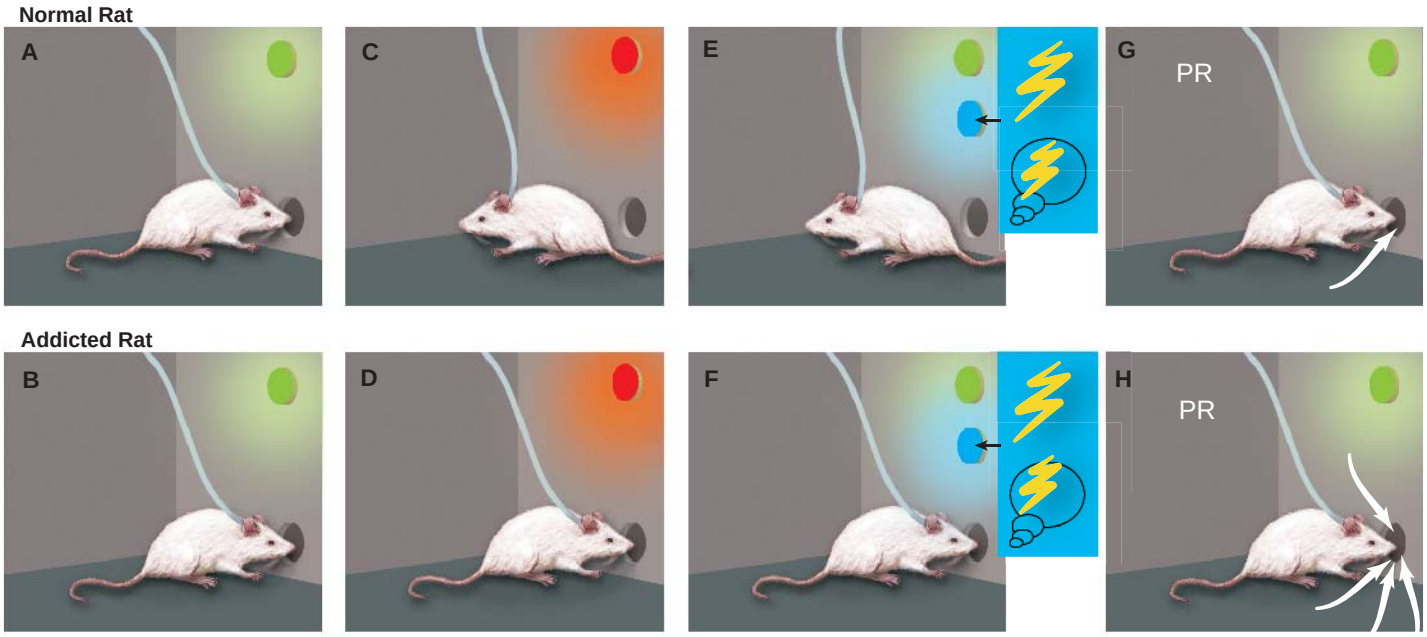
How do you tell whether a rat that has learned to self-administer a drug has become an "addict"? Mere self-administration is not evidence of addiction, because addiction refers to a specific pattern of compulsive drug-seeking and drug-taking behavior, one that predominates over most other activities in life. Indeed, most people have at some time self-administered a potentially addictive drug, but very few become addicts. What accounts for the transition from drug use to drug addiction, and why are some individuals more susceptible to this transition than others? Two papers on pages 1014 (*1*) and 1017 (*2*) of this issue represent a major advance in developing realistic preclinical animal models to answer these questions. Specifically, the two studies ask: How do you tell whether a rat has made the transition to addiction?

The author is in the Department of Psychology and Neuroscience Program, University of Michigan, Ann Arbor, MI 48109, USA. E-mail: ter@umich.edu

Nonhuman animals learn to avidly perform an action if it results immediately in the intravenous delivery of a potentially addictive drug, a phenomenon first reported in this journal by Weeks in 1962 (*3*). This self-administration animal model is still the "gold standard" for assessing the rewarding properties of drugs of abuse. From this model, we have learned a great deal about the conditions that support drug self-administration behavior. For example, nonhuman animals will self-administer nearly every drug that is self-administered by humans [with a few notable exceptions, such as hallucinogens (*4*)]. We also know that potentially addictive drugs usurp neural systems that evolved to mediate behaviors normally directed toward "natural rewards" [such as food, water, shelter, and sex (*5*)].

However, despite enormous advances, drug self-administration studies have not provided much insight into why some susceptible individuals undergo a transition to addiction, whereas others can maintain controlled drug use or forgo use altogether (*6*). This is in part because there have been no good animal models to distinguish mere drug self-administration behavior from the compulsive drug self-administration behavior that characterizes addiction. Deroche-Gamonet *et al*. (*1*) and Vanderschuren and Everitt (*2*) approached this problem in a straightforward yet elegant way. They identified three key diagnostic criteria for addiction and then simply asked whether rats allowed to self-administer cocaine for an extended period developed any of the symptoms of addiction described by the criteria.

The first diagnostic criterion selected is continued drug-seeking behavior even when the drug is known to be unavailable (*1*). This is reminiscent of the cocaine addict, who has run out of drug, compulsively searching the carpet for a few white crystals ("chasing ghosts") that they know will most likely be sugar. Deroche-Gamonet *et al*. (*1*) measured this behavior with two signals: a "go" cue that drug is available and a "stop" cue that drug is not available (see the figure). Normal rats quickly learn to work for drug only when the go cue is on, and refrain when the stop



When more is not enough. An innovative rat model for the study of addiction based on three diagnostic criteria (*1*, *2*). Shown are rat cages, each with a panel containing a hole through which a rat can poke its nose. Above the hole, a green light signals that cocaine is available. If the rat nose-pokes, it receives an intravenous injection of cocaine. (**A** and **B**) Under usual limited-access conditions, normal rats (A) and addicted rats (B) both self-administer cocaine at the same rate (*1*, *2*). If given a longer test session, however, addicted rats escalate their intake (*1*). (**C** and **D**) The red light indicates that cocaine is not available. Normal rats (C) stop responding, but addicted rats (D) continue to nose-poke even though cocaine is not delivered (*1*). (**E** and **F**) The green light signals that cocaine is available, but the additional blue light either indicates that cocaine delivery will be accompanied by a footshock (the lightning bolt) (*1*) or represents a cue previously associated with a footshock (the memory of shock) (*2*). Normal rats (E) decrease their responses in the presence of the blue light, but addicted rats (F) keep responding (*1*, *2*). (**G** and **H**) The green light signals that cocaine is available, but it is now available on a progressive ratio (PR) schedule where the number of responses required for an injection is progressively increased (for example, from 10 to 20, 30, 45, 65, 85, 155). Under these conditions, addicted rats (H) work harder than normal rats (G) for cocaine—that is, they show a higher "breaking point" (*1*).

cue is on. Addicted rats keep working even when signaled to stop.

The second criterion selected is unusually high motivation (desire) for the drug (*1*). A defining characteristic of addiction is a pathological desire ("craving") for the drug, which drives a willingness to exert great effort in its procurement. This criterion was measured with a progressive ratio schedule in which the amount of work required to obtain the drug progressively increased. At some point, the cost exceeds the benefit and animals stop working; this "breaking point" is thought to provide a measure of an animal's motivation to obtain a reward (*7*). Addicted rats have an increased breaking point (see the figure).

The final criterion is continued drug use even in the face of adverse consequences (*1*, *2*). Addicts often continue drug use despite dire consequences. This feature of addiction was modeled by asking whether rats would continue to work for cocaine even when their actions produced an electric shock along with the cocaine injection (*1*) or when the memory of past electric shocks was evoked (*2*). Addicted rats kept working despite negative consequences.

Of particular importance are the conditions under which these symptoms of addiction develop (which also explains why this demonstration has been so long in coming). These symptoms of addiction only appear after much more extensive drug self-administration experience than is the norm [see also (*8*)]. For the first month that animals self-administered cocaine, they did not show any symptoms. Only after more than a month of exposure to cocaine (*1*), or after sessions with prolonged drug access (*2*), did symptoms begin to emerge. Furthermore, Deroche-Gamonet *et al*. (*1*) report that after 3 months, only a small subset of animals became "addicts." Although they all avidly self-administered cocaine, 41% of rats failed to meet any of the three diagnostic criteria of addiction, 28% showed only one symptom, 14% two symptoms, and 17% all three symptoms. In addition, the animals that developed these symptoms were those that also showed a cardinal feature of addiction: a high propensity to relapse [as indicated by reinstatement of drug-seeking behavior elicited by either a drug "prime" or a drug-associated cue (*1*)]. Also of keen interest are measures not associated with these symptoms of addiction, including measures of anxiety, "impulsivity," and high versus low responsiveness to novelty (*1*). The researchers conclude that rats become "addicts" (i) only after extended experience with cocaine, and (ii) only if they are inherently susceptible.

Although extended access to cocaine led to continued drug-seeking in the face of adverse consequences in both studies, only Deroche-Gamonet *et al*. (*1*) found increased motivation for the drug. Vanderschuren and Everitt (*2*), however, used a very different and less traditional procedure for assessing motivation for drug, and their measure may be less sensitive (*7*). Consistent with the Deroche-Gamonet *et al*. findings (*1*), long daily sessions with continuous access to cocaine, which leads to escalation of intake (*9*), are associated with increased motivation for cocaine assessed using a progressive ratio schedule (*10*).

The demonstration that extended access to cocaine can lead to addiction-like behavior in the rat raises many questions. Would daily access to even more drug accelerate this process (*9*)? Does this happen with other addictive drugs? What differentiates susceptible from less susceptible individuals? Do less susceptible individuals become susceptible if given

CREDIT: TAINA LITWAK

more access to drug, or if exposed to, for example, stress or different environments? How does extended access to cocaine change the brain (and only in susceptible individuals) to produce different symptoms of addiction? In providing more realistic preclinical animal models of addiction than previously available, the two new reports set the stage for developing exciting new approaches with which to unravel the psychology and neurobiology of addiction.

References

1. V. Deroche-Gamonet, D. Belin, P. V. Piazza, *Science* **305**, 1014 (2004).
2. L. J. M. J. Vanderschuren, B. J. Everitt, *Science* **305**, 1017 (2004).
3. J. R. Weeks, *Science* **138**, 143 (1962).
4. R. A. Yokel, in *Methods of Assessing the Reinforcing Properties of Abused Drugs*, M. A. Bozarth, Ed. (Springer-Verlag, New York, 1987), pp. 1–33.
5. A. E. Kelley, K. C. Berridge, *J. Neurosci.* **22**, 3306 (2002).
6. T. E. Robinson, K. C. Berridge, *Annu. Rev. Psychol.* **54**, 25 (2003).
7. J. M. Arnold, D. C. Roberts, *Pharmacol. Biochem. Behav.* **57**, 441 (1997).
8. J. Wolffgramm, A. Heyne, *Behav. Brain Res.* **70**, 77 (1995).
9. S. H. Ahmed, G. F. Koob, *Science* **282**, 298 (1998).
10. N. E. Paterson, A. Markou, *Neuroreport* **14**, 2229 (2003).

CLIMATE SCIENCE

Already the Day After Tomorrow?

Bogi Hansen, Svein Østerhus, Detlef Quadfasel, William Turrell

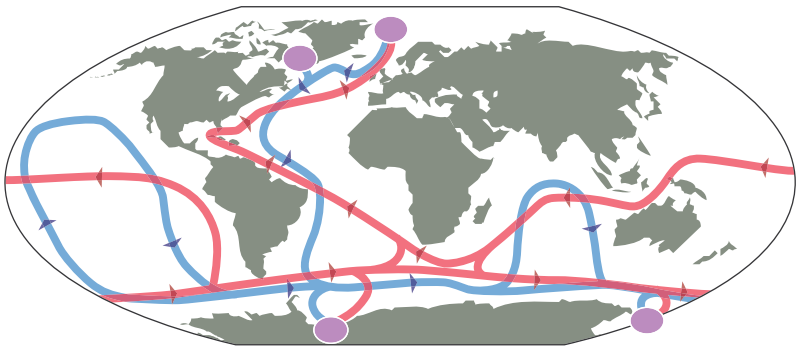
With even Hollywood aroused, the thermohaline circulation (THC) of the ocean has become a public theme, and not without reason. The THC helps drive the ocean currents around the globe and is important to the world's climate (see map on this page). There is a possibility that the North Atlantic THC may weaken substantially during this century, and this would have unpleasant effects on our climate—not a disaster-movie ice age, but perhaps a cooling over parts of northern Europe.

Enhanced online at www.sciencemag.org/cgi/content/full/305/5686/953

The THC is a driving mechanism for ocean currents. Cooling and ice formation at high latitudes increase the density of surface waters sufficiently to cause them to sink. Several different processes are involved, which collectively are termed "ventilation." When active, ventilation maintains a persistent supply of dense waters to the deep high-latitude oceans. At low latitudes, in contrast, vertical mixing heats the deep water and reduces its density. Together, high-latitude ventilation and low-latitude mixing build up horizontal density differences in the deep ocean, which generate forces. In the North Atlantic, these forces help drive the North Atlantic Deep Water (NADW) that supplies a large part of the deep waters of the world ocean.

Not everybody agrees that the THC is an important driving mechanism for the NADW flow. The north-south density differences observed at depth might be generated by the flow rather than driving it (*1*). This argument is tempting, but it neglects some salient features of the real ocean that are at odds with many conceptual, analytical, and even some numerical models.

The Greenland-Scotland Ridge splits the North Atlantic into two basins (see the figure on the next page). Most of the ventilation occurs in the northern basin, and the cold dense waters pass southward as deep overflows across the Ridge. According to measurements (*2*–*4*), the total volume transport across the Ridge attributable to these overflows is only about one-third of the total NADW production, but the volume transported approximately doubles by entrainment of ambient water within just a few hundreds of kilometers after passing the Ridge.



Thermohaline circulation. Schematic map of the thermohaline circulation of the world ocean. Purple ovals indicate ventilation areas, which feed the flow of deep dense waters (blue lines with arrows). These waters flow into all of the oceans and slowly ascend throughout them. From there, they return to the ventilation areas as warm compensating currents (red lines with arrows) in the upper layers.

On their way toward the Ridge, the overflow waters accelerate to current speeds of more than 1 m/s, which is clear evidence of THC forcing. After crossing the Ridge, the flows descend to great depths in bottom currents, which again are density-driven. In the present-day ocean, THC drives the overflows, which together with the entrained water feed most of the NADW.

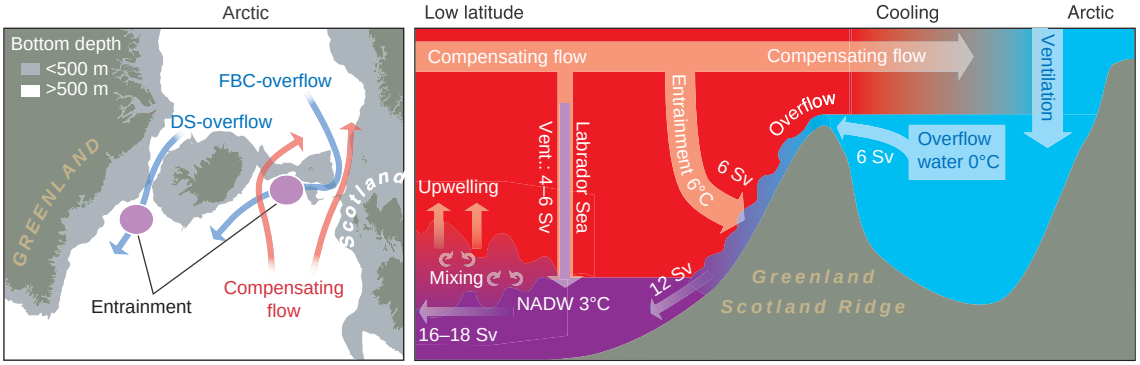
This is the reason why people worry about a possible weakening of the THC. In the coming decades, global change via atmospheric pathways is expected to increase the freshwater supply to the Arctic. This will reduce the salinity and hence the density of surface waters, and thereby may reduce ventilation. Even if the ventilation comes to a total halt, this will not stop the overflows immediately, because the reservoir of dense water north of the Ridge stabilizes the overflow. Instead, the supply of NADW would diminish in a matter of decades. In contrast, large changes in low-latitude mixing—even if conceivable—require a much longer time before affecting the THC (*5*).

A potential weakening of the North Atlantic THC would affect the deep waters of the world ocean in the long run, but would have more immediate effects on the climate in some regions. The dense overflow waters feeding the deep Atlantic are replenished by a compensating northward flow in the upper layers. These currents bring warm saline water northward to the regions where ventilation and entrainment occur. This oceanic heat transport keeps large Arctic areas free of ice and parts of the North Atlantic several degrees warmer than they would otherwise have been (*6*).

A substantially weakened THC reduces this heat transport and regionally counterbalances global warming. In some areas, it might even lead to cooling (*7*). This has in-

B. Hansen is at the Faroese Fisheries Laboratory, FO-110 Torshavn, Faroe Islands. S. Østerhus is at the Bjerknes Center, NO-5007 Bergen, Norway. D. Quadfasel is at the Institut für Meereskunde, D-20146 Hamburg, Germany. W. Turrell is at the Marine Laboratory, Aberdeen AB11 9DB, Scotland. E-mail: bogihan@frs.fo

spired a public debate focused on a potential cooling of northern Europe, which has the compensating flow just off the coast. Note that this part of the North Atlantic THC is especially dependent on ventilation north of the Greenland-Scotland Ridge, overflow, and entrainment (*3*).



North Atlantic flow. The exchange of water across the Greenland-Scotland Ridge is a fundamental component of the North Atlantic THC. Arrows on the map indicate the main overflow (blue) and compensating inflow (red) branches. On the schematic section to the right, temperatures in °C and volume transports in Sv (1 Sv = 10^6 m^3/s) are approximate values. DS, Denmark Strait; FBC, Faroe Bank Channel.

The concept of a weakened THC is supported by some numerical climate models (*8*), but not by all. Increased salinity of the compensating flow may balance the salinity decrease from the increased freshwater supply and maintain ventilation (*9*). Climate models, so far, do not provide a unique answer describing the future development of the THC, but what is the present observational evidence?

It is argued that early evidence for changes should primarily be sought in the ventilation and overflow rates. Indeed, some such changes have been reported. Since around 1960, large parts of the open sea areas north of the Greenland-Scotland Ridge have freshened (*10*), and so have the overflows (*11*). At the same time, low-latitude Atlantic waters became more saline in the upper layer (*12*), and this is also reflected in the compensating flow. Long-term observations in both of the main branches of compensating flow across the Greenland-Scotland Ridge have shown increasing salinity since the mid-1970s, with a record high in 2003.

Even more convincing evidence for a reduction of the North Atlantic THC has been gained from monitoring both the overflows and the compensating northward flow by direct current measurements (*13*). For the Denmark Strait overflow, no persistent long-term trends in volume transport have been reported (*2*, *14*), but the Faroe Bank Channel overflow was found to have decreased by about 20% from 1950 to 2000 (*15*).

We find evidence of freshening of the Nordic Seas and a reduction of the strength of the overflow, both of which will tend to weaken the North Atlantic THC. On the other hand, the compensating northward flow is getting more saline, which may maintain ventilation and counterbalance the THC decrease. So the jury is still out. This emphasizes the need for more refined climate models and long-term observational systems that are capable of identifying potential changes in our climate system.

References

1. C. Wunsch, *Science* **298**, 1179 (2002).
2. R. R. Dickson, J. Brown, *J. Geophys. Res.* **99**, 12,319 (1994).
3. B. Hansen, S. Østerhus, *Prog. Oceanogr.* **45**,109 (2000).
4. A. Ganachaud, C. Wunsch, *Nature* **408**, 453 (2000).
5. W. Munk, C. Wunsch, *Deep-Sea Res.* **45**, 1976 (1998).
6. R. Seager *et al.*, *Q. J. R. Meteorol. Soc.* **128**, 2563 (2002).
7. M. Vellinga, R. A. Wood, *Clim. Change* **54**, 251 (2002).
8. S. Rahmstorf, *Nature* **399**, 523 (1999).
9. M. Latif *et al.*, *J. Clim.* **13**, 1809 (2000).
10. J. Blindheim *et al.*, *Deep-Sea Res. I* **47**, 655 (2000).
11. R. R. Dickson *et al.*, *Nature* **416**, 832 (2002).
12. R. Curry *et al.*, *Nature* **426**, 826 (2003).
13. Arctic/Subarctic Ocean Fluxes (ASOF) (http://asof.npolar.no).
14. R. R. Dickson, personal communication.
15. B. Hansen, W. R. Turrell, S. Østerhus, *Nature* **411**, 927 (2001).

NEUROSCIENCE

NAD to the Rescue

Antonio Bedalov and Julian A. Simon

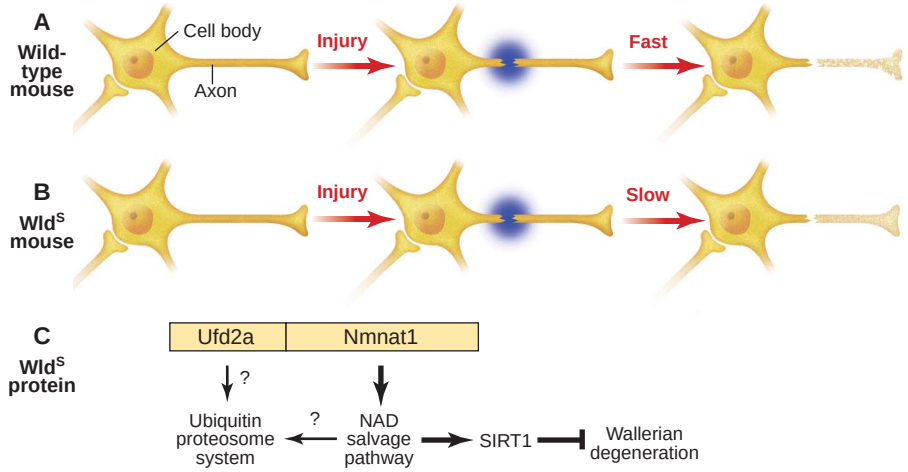
The cofactor nicotinamide adenine dinucleotide (NAD)—once consigned to the oblivion of metabolic pathway wall charts—has recently attained celebrity status as the link between metabolic activity, cellular resistance to stress or injury, and longevity. NAD influences many cell fate decisions—for example, NAD-dependent enzymes such as poly (ADP-ribose) polymerase (PARP) are important for the DNA damage response, and NAD-dependent protein deacetylases (Sirtuins) are involved in transcriptional regulation, the stress response, and cellular differentiation. On page 1010 of this issue, Araki and colleagues (*1*) extend the influence of NAD with their demonstration that an increase in NAD biosynthesis or enhanced activity of the NAD-dependent deacetylase SIRT1 protects mouse neurons from mechanical or chemical injury (*2*).

Axonal degeneration (termed Wallerian degeneration) often precedes the death of neuronal cell bodies in neurodegenerative diseases such as Alzheimer's (AD) and Parkinson's (PD). Mice carrying the spontaneous dominant *Wld*^{s} mutation show delayed axonal degeneration following neuronal injury. The *Wld*^{s} mutation on mouse chromosome 4 is a rare tandem triplication of an 85-kb DNA fragment that harbors a translocation. The translocation encodes a fusion protein comprising the amino-terminal 70 amino acids of Ufd2a (ubiquitin fusion degradation protein 2a), an E4 ubiquitin ligase, and the entire coding region of Nmnat1 (nicotinamide mononucleotide adenylyltranferase 1), an NAD biosynthetic enzyme. Although the C57BL/*Wld*^{s} mouse was described 15 years ago (*3*) and expression of the Wld^s fusion protein is known to delay Wallerian degeneration (*4*), the mechanism of neuroprotection has remained elusive. Given that proteasome inhibitors block Wallerian degeneration both in vitro and in vivo (*5*), the Ufd2a protein fragment (a component of the ubiquitin proteasome system) has been the prime candidate for mediator of neuroprotection in the *Wld*^{s} mouse. Indeed, ubiquitin-mediated protein degradation by the proteasome

A. Bedalov is in the Clinical Research Division and J. A. Simon is in the Clinical Research and Human Biology Divisions, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. E-mail: abedalov@fhcrc.org, jsimon@fhcrc.org

has been identified as a potential target for developing drugs to treat neurodegenerative diseases such as AD, PD, and multiple sclerosis (*6, 7*).

Araki *et al*. (*1*) developed an in vitro model of Wallerian degeneration comprising cultures of primary dorsal root ganglion neurons derived from wild-type mice. The neurons overexpressed either the Wld^s fusion protein or one of the fusion protein fragments. Surprisingly, the authors found that overexpression of the Ufd2a protein fragment alone did not delay degeneration of axons injured by removal of the neuronal cell body (transection) or treatment with the neurotoxin vincristine. In contrast, overexpression of Nmnat1 or the addition of NAD to the neuronal cultures before injury delayed axonal degeneration in response to mechanical or chemical damage.



Energizing neuroprotection. (**A**) In wild-type mice, axons of injured neurons rapidly degenerate (Wallerian degeneration) in a process that may be relevant to the neurodegeneration seen in diseases like AD and PD. (**B**) In mice with the *Wld^s* dominant mutation (a tandem triplication of a region on mouse chromosome 4), injured neurons show a delay in Wallerian degeneration due to activity of the Wld^s fusion protein. (**C**) The fusion protein consists of the amino terminus of Ufd2a (an E4 ubiquitin-conjugating enzyme) and the entire sequence of Nmnat1 (an enzyme in the NAD salvage pathway). Neuroprotection in the *Wld^s* mouse may result from increased synthesis of NAD, leading to a concomitant increase in the activity of the NAD-dependent deacetylase, SIRT1, which may activate a transcription factor that induces expression of genes involved in neuroprotection (*1*).

It is well established that increased expression of NAD salvage pathway genes in yeast, including the yeast homologs of Nmnat1 (*NMA1* and *NMA2*), lengthens life-span and boosts resistance to stress, an effect that depends on the NAD-dependent deacetylase Sir2 (*8*). Based on this observation, Araki *et al*. tested whether the protective effect of increased Nmnat1 expression required NAD-dependent deacetylase activity. Expression of small interfering RNAs that target each of the seven Sir2 mammalian homologs (SIRT1 through SIRT7) decreased survival of the dorsal root ganglion cultures after injury only when SIRT1 expression was reduced. The same effect was observed when SIRT1 activity was blocked with a small-molecule inhibitor; a SIRT1 activator, on the other hand, boosted neuronal survival following injury. These data suggest that protection against Wallerian degeneration is the result of increased expression of Nmnat1, a rise in nuclear NAD levels, and a consequent increase in SIRT1 activity. This conclusion does not negate the involvement of the proteasome in Wallerian degeneration, but it does indicate that the protective effect of the Wld^s fusion protein is independent of Ufd2a activity. Indeed, the new findings throw open the possibility that changes in NAD levels may indirectly regulate the ubiquitin-proteasome system.

The enzymes SIRT1 through SIRT7 belong to a unique enzyme class that requires a boost in NAD levels to maintain activity, because they consume this cofactor during deacetylation of target proteins. Another enzyme that depletes cellular NAD levels is PARP. In the presence of NAD, inhibition of PARP has little effect on Wallerian degeneration; however, in the absence of exogenous NAD, inhibition of PARP increases the survival of dorsal root ganglion cultures after injury (*1*). This suggests that neuronal survival requires the maintenance of adequate NAD levels, but that a boost in NAD levels beyond this point confers no additional benefit.

In intact neurons of C57BL/*Wld^s* mice, the Wld^s fusion protein is expressed almost exclusively in the nucleus (*4*). In fibroblasts (*9*)—and, presumably, in neurons—SIRT1 also is expressed in the nucleus. SIRT1 and other NAD-dependent deacetylases alter gene expression by targeting histone proteins as well as key nuclear transcription factors such as p53 (*9*, *10*), forkhead (*11*, *12*), and NF-κB (*13*). In addition, Sirtuins also deacetylate cytoplasmic proteins, including α-tubulin. The protective effect of the Wld^s fusion protein appears to be exerted in the nucleus, because addition of NAD after removal of cell bodies in the neuronal cultures is no longer protective. This suggests that an alternative program of gene expression is initiated by elevated NAD levels in the nucleus, leading to the production of protective factors that actively block Wallerian degeneration. The therapeutic implication of this finding is that it may be possible to design neuroprotective drugs that boost SIRT1 activity and prevent further neurodegeneration in diseases like AD and PD.

The Araki *et al*. study (*1*) addresses the long-standing question of how the Wld^s fusion protein prevents Wallerian degeneration. As with most groundbreaking studies, new questions emerge. For example, what is the direct result of increased Nmnat1 expression? Overexpression of Nmnat1 leads to increased activity of this enzyme but does not change total NAD levels or the ratio of NAD to NADH, raising the possibility that increased Nmnat1 activity may result in a decrease in nicotinamide or other inhibitory molecules. It is possible that the relevant target of SIRT1's neuroprotective activity may be a transcription factor that responds to changes in the cell's metabolic state by switching on expression of genes that encode neuroprotective proteins. Identifying the targets of SIRT1 that mediate the neuroprotective effect may broaden the options for therapeutic intervention in AD, PD, and other neurodegenerative diseases.

References

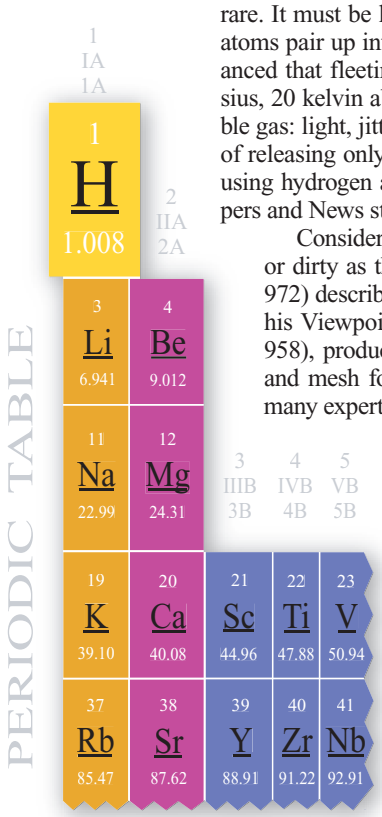
1. T. Araki, Y. Sasaki, J. Milbrandt, *Science* **305**, 1010 (2004).
2. A. Waller, *Philos. Trans. R Soc. London* **140**, 423 (1850).
3. E. R. Lunn *et al*., *Eur. J. Neurosci.* **1**, 27 (1989).
4. T. G. Mack *et al*., *Nature Neurosci.* **4**, 1199 (2001).
5. Q. Zhai *et al*., *Neuron*, **39**, 217 (2003).
6. M. P. Coleman, V. H. Perry, *Trends Neurosci.* **25**, 532 (2002).
7. A. Sadaji, B. L. Schneider, P. Aebischer, *Curr. Biol.* **14**, 326 (2004).
8. R. M. Anderson *et al*., *J. Biol. Chem.* **277**, 18881 (2002).
9. H. Vaziri *et al*., *Cell* **107**, 149 (2001).
10. J. Luo *et al*., *Cell* **107**, 137 (2001).
11. A. Brunet *et al*., *Science* **303**, 2011 (2004).
12. M. C. Motta *et al*., *Cell* **116**, 551 (2004).
13. F. Yeung *et al*., *EMBO J.* **23**, 2369 (2004).

INTRODUCTION

Not So Simple

Seen up close, hydrogen looks like a recipe for success. Small and simple—one proton and one electron in its most common atomic form—hydrogen was the first element to assemble as the universe cooled off after the big bang, and it is still the most widespread. It accounts for 90% of the atoms in the universe, two-thirds of the atoms in water, and a fair proportion of the atoms in living organisms and their geologic legacy, fossil fuels.

To scientists and engineers, those atoms offer both promise and frustration. Highly electronegative, they are eager to bond, and they release energy generously when they do. That makes them potentially useful, if you can find them. On Earth, however, unattached hydrogen is vanishingly rare. It must be liberated by breaking chemical bonds, which requires energy. Once released, the atoms pair up into two-atom molecules, whose dumbbell-shaped electron clouds are so well balanced that fleeting charge differences can pull them into a liquid only at a frigid –252.89° Celsius, 20 kelvin above absolute zero. The result, at normal human-scale temperatures, is an invisible gas: light, jittery, and slippery; hard to store, transport, liquefy, and handle safely; and capable of releasing only as much energy as human beings first pump into it. All of which indicates that using hydrogen as a common currency for an energy economy will be far from simple. The papers and News stories in this special section explore some of its many facets.



Consider hydrogen's green image. As a manufactured product, hydrogen is only as clean or dirty as the processes that produce it in the first place. Turner (p. 972) describes various options for large-scale hydrogen production in his Viewpoint. Furthermore, as News writer Service points out (p. 958), production is just one of many technologies that must mature and mesh for hydrogen power to become a reality, a fact that leads many experts to urge policymakers to cast as wide a net as possible.

In some places, the transition to hydrogen may be relatively straightforward. For her News story (p. 966), Vogel visited Iceland, whose abundant natural energy resources have given it a clear head start. Elsewhere, though, various technological detours and bridges may lie ahead. The Viewpoint by Demirdöven and Deutch (p. 974) and Cho's News story (p. 964) describe different intermediate technologies that may shape the next generation of automobiles. Meanwhile, the fires of the fossil fuel–based "carbon economy" seem sure to burn intensely for at least another half-century or so [see the Editorial by Kennedy (p. 917)]. Service's News story on carbon sequestration (p. 962) and Pacala and Socolow's Review (p. 968) explore strategies—including using hydrogen—for mitigating their effects.

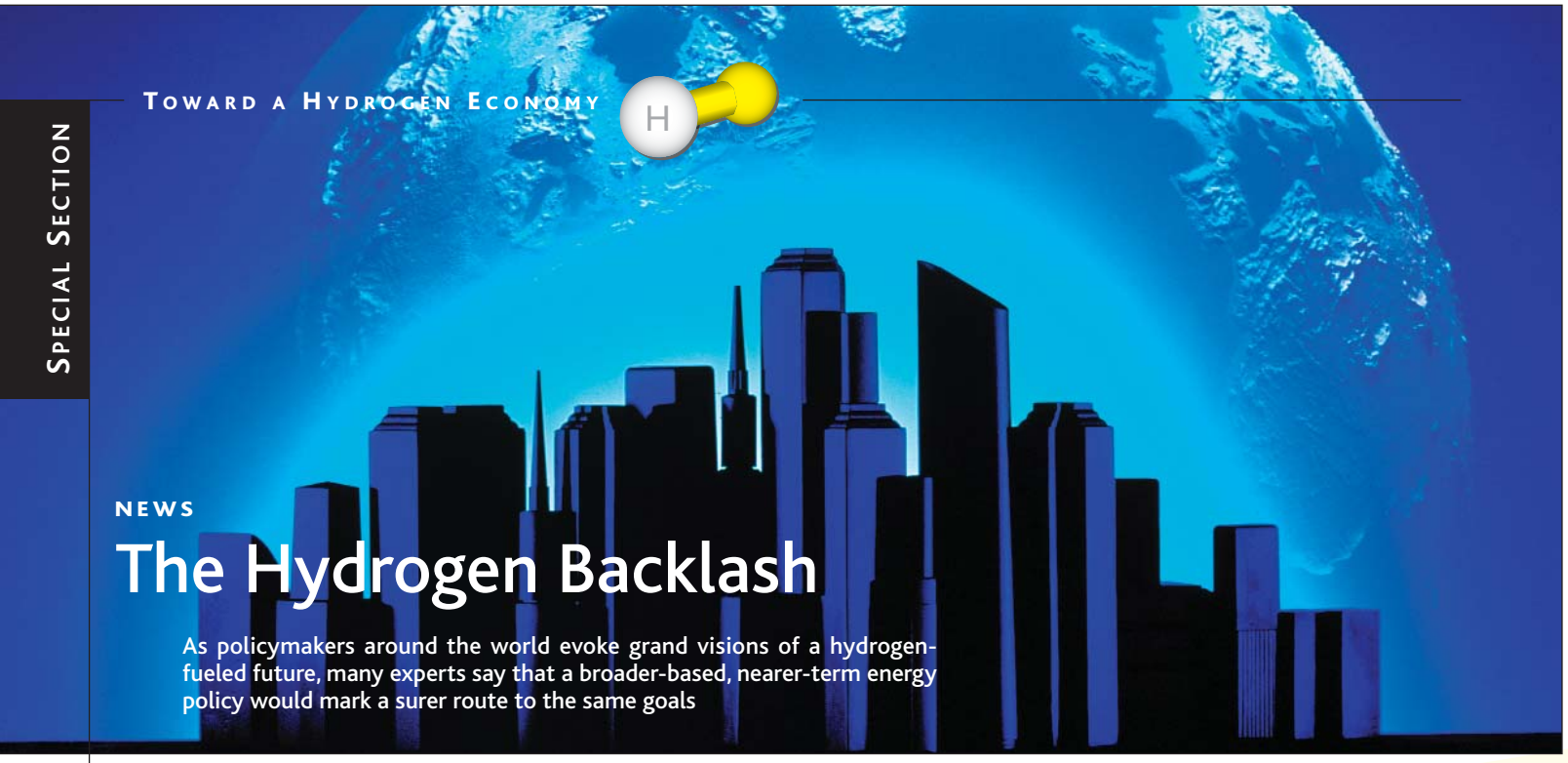
Two generations down the line, the world may end up with a hydrogen economy completely different from the one it expected to develop. Perhaps the intermediate steps on the road to hydrogen will turn out to be the destination. The title we chose for this issue—Toward a Hydrogen Economy—reflects that basic uncertainty and the complexity of what is sure to be a long, scientifically engaging journey.

–ROBERT COONTZ AND BROOKS HANSON

CONTENTS

See also related Editorial on p. 917.

Science



NEWS

The Hydrogen Backlash

As policymakers around the world evoke grand visions of a hydrogen-fueled future, many experts say that a broader-based, nearer-term energy policy would mark a surer route to the same goals

In the glare of a July afternoon, the HydroGen3 minivan threaded through the streets near Capitol Hill. As a *Science* staffer put it through its stop-and-go paces, 200 fuel cells under the hood of the General Motors prototype inhaled hydrogen molecules, stripped off their electrons, and fed current to the electric engine. The only emissions: a little extra heat and humidity. The result was a smooth, eerily quiet ride—one that, with H3's priced at $1 million each, working journalists won't be repeating at their own expense anytime soon.

Hydrogen-powered vehicles may be rareties on Pennsylvania Avenue, but in Washington, D.C., and other world capitals they and their technological kin are very much on people's minds. Switching from fossil fuels to hydrogen could dramatically reduce urban air pollution, lower dependence on foreign oil, and reduce the buildup of greenhouse gases that threaten to trigger severe climate change.

With those perceived benefits in view, the United States, the European Union, Japan, and other governments have sunk billions of dollars into hydrogen initiatives aimed at revving up the technology and propelling it to market. Car and energy companies are pumping billions more into building demonstration fleets and hydrogen fueling stations. Many policymakers see the move from oil to hydrogen as manifest destiny, challenging but inevitable. In a recent speech, Spencer Abraham, the U.S. secretary of energy, said such a transformation has "the potential to change our country on a scale of the development of electricity and the internal combustion engine."

The only problem is that the bet on the hydrogen economy is at best a long shot. Recent reports from the U.S. National Academy of Sciences (NAS) and the American Physical Society (APS) conclude that researchers face daunting challenges in finding ways to produce and store hydrogen, convert it to electricity, supply it to consumers, and overcome vexing safety concerns. Any of those hurdles could block a broad-based changeover. Solving them simultaneously is "a very tall order," says Mildred Dresselhaus, a physicist at the Massachusetts Institute of Technology (MIT), who has served on recent hydrogen review panels with the U.S. Department of Energy (DOE) and APS as well as serving as a reviewer for the related NAS report.

As a result, the transition to a hydrogen economy, if it comes at all, won't happen soon. "It's very, very far away from substantial deployed impact," says Ernest Moniz, a physicist at MIT and a former undersecretary of energy at DOE. "Let's just say decades, and I don't mean one or two."

In the meantime, some energy researchers complain that, by skewing research toward costly large-scale demonstrations of technology well before it's ready for market, governments risk repeating a pattern that has sunk previous technologies such as synfuels in the 1980s. By focusing research on technologies that aren't likely to have a measurable impact until the second half of the century, the current hydrogen push fails to address the growing threat from greenhouse gas emissions from fossil fuels. "There is starting to be some backlash on the hydrogen economy," says Howard Herzog, an MIT chemical engineer. "The hype has been way overblown. It's just not thought through."

A perfect choice?

Almost everyone agrees that producing a viable hydrogen economy is a worthy long-term goal. For starters, worldwide oil production is expected to peak within the next few decades, and although supplies will remain plentiful long afterward, oil prices are expected to soar as international markets view the fuel as increasingly scarce. Natural gas production is likely to peak a couple of decades after oil. Coal, tar sands, and other fossil fuels should remain plentiful for at least another century. But these dirtier fuels carry a steep environmental cost: Generating electricity from coal instead of natural gas, for example, releases twice as much carbon dioxide (CO_2). And in order to power vehicles, they must be

CREDIT: (TOP TO BOTTOM) GEORGE B. DIEBOLD/CORBIS; PHOTO COURTESY OF GM CORP.

converted to a liquid or gas, which requires energy and therefore raises their cost.

Even with plenty of fossil fuels available, it's doubtful we'll want to use them all. Burning fossil fuels has already increased the concentration of CO_2 in the atmosphere from 280 to 370 parts per million (ppm) over the past 150 years. Unchecked, it's expected to pass 550 ppm this century, according to New York University physicist Martin Hoffert and colleagues in a 2002 *Science* paper (*Science*, 1 November 2002, p. 981). "If sustained, [it] could eventually produce global warming comparable in magnitude but opposite in sign to the global cooling of the last Ice Age," the authors write. Development and population growth can only aggravate the problems.

On the face of it, hydrogen seems like the perfect alternative. When burned, or oxidized in a fuel cell, it emits no pollution, including no greenhouse gases. Gram for gram, it releases more energy than any other fuel. And as a constituent of water, hydrogen is all around us. No wonder it's being touted as the clean fuel of the future and the answer to modern society's addiction to fossil fuels. In April 2003, *Wired* magazine laid out "How Hydrogen Can Save America." Environmental gadfly Jeremy Rifkin has hailed the hydrogen economy as the next great economic revolution. And General Motors has announced plans to be the first company to sell 1 million hydrogen fuel cell cars by the middle of the next decade.

Last year, the Bush Administration plunged in, launching a 5-year, $1.7 billion initiative to commercialize hydrogen-powered cars by 2020. In March, the European Commission launched the first phase of an expected 10-year, €2.8 billion public-private partnership to develop hydrogen fuel cells. Last year, the Japanese government nearly doubled its fuel cell R&D budget to $268 million. Canada, China, and other countries have mounted efforts of their own. Car companies have already spent billions of dollars trying to reinvent their wheels—or at least their engines—to run on hydrogen: They've turned out nearly 70 prototype cars and trucks as well as dozens of buses. Energy and car companies have added scores of hydrogen fueling stations worldwide, with many more on the drawing boards (see p. 964). And the effort is still gaining steam.

The problem of price

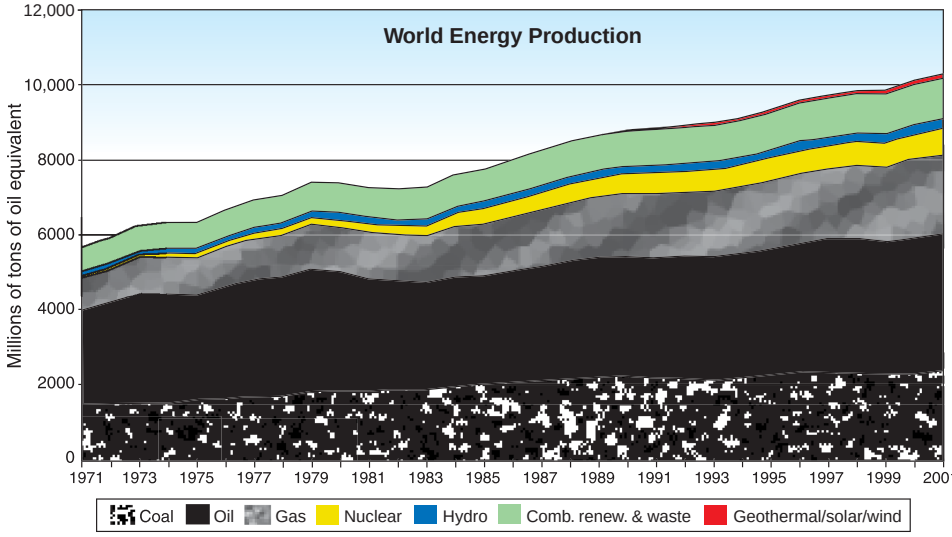
Still, despite worthwhile goals and good intentions, many researchers and energy experts say current hydrogen programs fall pitifully short of what's needed to bring a hydrogen economy to pass. The world's energy infrastructure is too vast, they say, and the challenges of making hydrogen technology competitive with fossil fuels too daunting unless substantially more funds are added to the pot. The current initiatives are just "a start," Dresselhaus says. "None of the reports say it's impossible," she adds. However, Dresselhaus says, "the problem is very difficult no matter how you slice it."

Economic and political difficulties abound, but the most glaring barriers are technical. At the top of the list: finding a simple and cheap way to produce hydrogen. As is often pointed out, hydrogen is not a fuel in itself, as oil and coal are. Rather, like electricity, it's an energy carrier that must be generated using another source of power. Hydrogen is the most common element in the universe. But on Earth, nearly all of it is bound to other elements in molecules, such as hydrocarbons and water. Hydrogen atoms must be split off these molecules to generate dihydrogen gas (H_2), the form it needs to be in to work in most fuel cells. These devices then combine hydrogen and oxygen to make water and liberate electricity in the process. But every time a fuel is converted from one source, such as oil, to another, such as electricity or hydrogen, it costs energy and therefore money.

Today, by far the cheapest way to produce hydrogen is by using steam and catalysts to break down natural gas into H_2 and CO_2. But although the technology has been around for decades, current steam reformers are only 85% efficient, meaning that 15% of the energy in natural gas is lost as waste heat during the reforming process. The upshot, according to Peter Devlin, who runs a hydrogen production program at DOE, is that it costs $5 to produce the amount of hydrogen that releases as much energy as a gallon of gasoline. Current techniques for liberating hydrogen from coal, oil, or water are even less efficient. Renewable energy such as solar and wind power can also supply electricity to split water, without generating CO_2. But those technologies are even more expensive. Generating electricity with solar power, for example, remains 10 times more expensive than doing so with a coal plant. "The energy in hydrogen will always be more expensive than the sources used to make it," said Donald Huberts, chief executive officer of Shell Hydrogen, at a hearing before the U.S. House Science Committee in March. "It will be competitive only by its other benefits: cleaner air, lower greenhouse gases, et cetera."

The good news, Devlin says, is that production costs have been coming down, dropping about $1 per gallon ($0.25/liter) of gasoline equivalent over the past 3 years. The trouble is that DOE's own road map projects that drivers will buy hydrogen-powered cars only if the cost of the fuel drops to $1.50 per gallon of gasoline equivalent by 2010 and even lower in the years beyond. "The easy stuff is over," says Devlin. "There are going to have to be some fundamental breakthroughs to get to $1.50."

There are ideas on the drawing board. In addition to stripping hydrogen from fossil fuels, DOE and other funding agencies are backing innovative research ideas to produce hydrogen with algae, use sunlight and catalysts to split water molecules directly, and siphon hydrogen from agricultural waste and other types of "biomass." Years of research in all of these areas, however, have yet to yield decisive progress.



Over a barrel. The world is growing increasingly dependent on fossil fuels.

SOURCE: INTERNATIONAL ENERGY AGENCY

To have and to hold

If producing hydrogen cheaply has researchers scratching their heads, storing enough of it on board a car has them positively stymied. Because hydrogen is the lightest element, far less of it can fit into a given volume than other fuels. At room temperature and pressure, hydrogen takes up roughly 3000 times as much space as gasoline containing the same amount of energy. That means storing enough of it in a fuel tank to drive 300 miles (483 kilometers)—DOE's benchmark—requires either compressing it, liquefying it, or using some other form of advanced storage system.

Unfortunately, none of these solutions is up to the task of carrying a vehicle 300 miles on a tank. Nearly all of today's prototype hydrogen vehicles use compressed gas. But these are still bulky. Tanks pressurized to 10,000 pounds per square inch (70 MPa) take up to eight times the volume of a current gas tank to store the equivalent amount of fuel. Because fuel cells are twice as efficient as gasoline internal combustion engines, they need fuel tanks four times as large to propel a car the same distance.

Liquid hydrogen takes up much less room but poses other problems. The gas liquefies at –253°C, just a few degrees above absolute zero. Chilling it to that temperature requires about 30% of the energy in the hydrogen. And the heavily insulated tanks needed to keep liquid fuel from boiling away are still larger than ordinary gasoline tanks.

Other advanced materials are also being investigated to store hydrogen, such as carbon nanotubes, metal hydrides, and substances such as sodium borohydride that produce hydrogen by means of a chemical reaction. Each material has shown some promise. But for now, each still has fatal drawbacks, such as requiring high temperature or pressures, releasing the hydrogen too slowly, or requiring complex and time-consuming materials recycling. As a result, many experts are pessimistic. A report last year from DOE's Basic Energy Sciences Advisory Committee concluded: "A new paradigm is required for the development of hydrogen storage materials to facilitate a hydrogen economy." Peter Eisenberger, vice provost of Columbia University's Earth Institute, who chaired the APS report, is even more blunt. "Hydrogen storage is a potential showstopper," he says.

Breakthroughs needed

Another area in need of serious progress is the fuel cells that convert hydrogen to electricity. Fuel cells have been around since the 1800s and have been used successfully for decades to power spacecraft. But their high cost and other drawbacks have kept them from being used for everyday applications such as cars. Internal combustion engines typically cost $30 for each kilowatt of power they produce. Fuel cells, which are loaded with precious-metal catalysts, are 100 times more expensive than that.

If progress on renewable technologies is any indication, near-term prospects for cheap fuel cells aren't bright, says Joseph Romm, former acting assistant secretary of energy for renewable energy in the Clinton Administration and author of a recent book, *The Hype About Hydrogen: Fact and Fiction in the Race to Save the Climate*. "It has taken wind power and solar power each about twenty years to see a tenfold decline in prices, after major government and private sector investments, and they still each comprise well under 1% of U.S. electricity generation," Romm said in written testimony in March before the House Science Committee reviewing the Administration's hydrogen initiative. "A major technology breakthrough is needed in transportation fuel cells before they will be practical." Various technical challenges—such as making fuel cells rugged enough to withstand the shocks of driving and ensuring the safety of cars loaded with flammable hydrogen gas—are also likely to make hydrogen cars costlier to engineer and slower to win public acceptance.

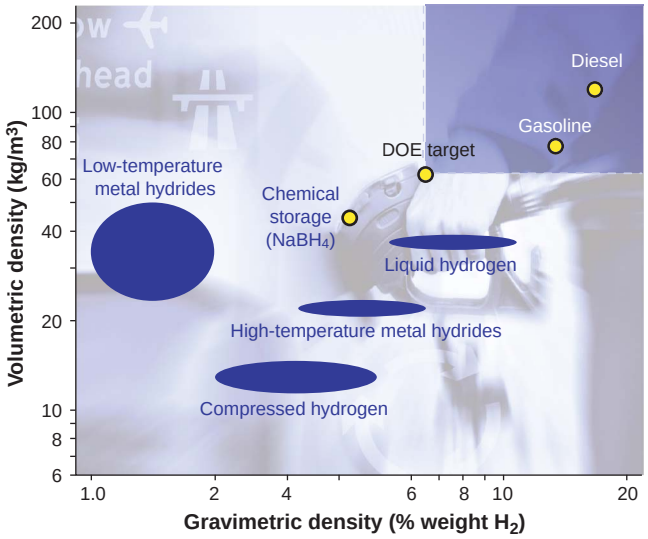
If they clear their internal technical hurdles, hydrogen fuel cell cars face an obstacle from outside: the infrastructure they need to refuel. If hydrogen is generated in centralized plants, it will have to be trucked or piped to its final destination. But because of hydrogen's low density, it would take 21 tanker trucks to haul the amount of energy a single gasoline truck delivers today, according to a study by Switzerland-based energy researchers Baldur Eliasson and Ulf Bossel. A hydrogen tanker traveling 500 kilometers would devour the equivalent of 40% of its cargo.

Ship the hydrogen as a liquid? Commercial-scale coolers are too energy-intensive for the job, Eliasson and Bossel point out. Transporting hydrogen through long-distance pipelines wouldn't improve matters much. Eliasson and Bossel calculate that 1.4% of the hydrogen flowing through a pipeline would be required to power the compressors needed to pump it for every 150 kilometers the gas must travel. The upshot, Eliasson and Bossel report: "Only 60% to 70% of the hydrogen fed into a pipeline in Northern Africa would actually arrive in Europe."

To lower those energy penalties, some analysts favor making hydrogen at fueling stations or in homes where it will be used, with equipment powered by the existing electricity grid or natural gas. But onsite production wouldn't be cheap, either. Eliasson and Bossel calculate that to supply hydrogen for 100 to 2000 cars per day, an electrolysis-based fueling station would require between 5 and 81 megawatts of electricity. "The generation of hydrogen at filling stations would make a threefold increase of electric power generating capacity necessary," they report. And at least for the foreseeable future, that extra electricity is likely to come from fossil fuels.

Whichever approach wins out, it will need a massive new hydrogen infrastructure to deliver the goods. The 9 million tons of hydrogen (enough to power between 20 million and 30 million cars) that the United States produces yearly for use in gasoline refining and chemical plants pale beside the needs of a full-blown transportation sector. For a hydrogen economy to catch on, the fuel must be available in 30% to 50% of filling stations when mass-market hydrogen cars become available, says Bernard Bulkin, former chief scientist at BP. A recent study by Marianne Mintz and colleagues at Argonne National Laboratory in Illinois found that creating the infrastructure needed to fuel 40% of America's cars would cost a staggering $500 billion or more.

Energy and car companies are unlikely



Showstopper? Current hydrogen storage technologies fall short of both the U.S. Department of Energy target and the performance of petroleum.

CREDIT: ADAPTED FROM P. GUAY, *MODÉLISATION MONTE-CARLO DE L'ADSORPTION DE L'HYDROGÈNE DANS LES NANOSTRUCTURES DE CARBONE*, INRS-EMT, MONTRÉAL, 2003; (IMAGE) ROYALTY-FREE/CORBIS

to spend such sums unless they know mass-produced hydrogen vehicles are on the way. Carmakers, however, are unlikely to build fleets of hydrogen vehicles without stations to refuel them. "We face a 'chicken and egg' problem that will be difficult to overcome," said Michael Ramage, a former executive vice president of ExxonMobil Research and Engineering, who chaired the NAS hydrogen report, when the report was released in February.

Stress test

Each of the problems faced by the hydrogen economy—production, storage, fuel cells, safety, and infrastructure—would be thorny enough on its own. For a hydrogen economy to succeed, however, all of these challenges must be solved simultaneously. One loose end and the entire enterprise could unravel. Because many of the solutions require fundamental breakthroughs, many U.S. researchers question their country's early heavy emphasis on expensive demonstration projects of fuel cell cars, fueling stations, and other technologies.

CO_2 free. To be a clean energy technology, hydrogen must be generated from wind, solar, or other carbon-free sources.

To illustrate the dangers of that approach, the APS report cites the fate of synfuels research in the 1970s and '80s. President Gerald Ford proposed that effort in 1975 as a response to the oil crisis of the early 1970s. But declining oil prices in the 1980s and unmet expectations from demonstration projects undermined industrial and congressional support for the technology. For hydrogen, the report's authors say, the "enormous performance gaps" between existing technology and what is needed for a hydrogen economy to take root means that "the program needs substantially greater emphasis on solving the fundamental science problems."

Focusing the hydrogen program on basic research will naturally give it the appropriate long-term focus it deserves, Romm and others believe. In the meantime, they say, the focus should be on slowing the buildup of greenhouse gases. "If we fail to limit greenhouse gas emissions over the next decade—and especially if we fail to do so because we have bought into the hype about hydrogen's near-term prospects—we will be making an unforgivable national blunder that may lock in global warming for the U.S. of 1 degree Fahrenheit [0.56°C] per decade by mid-century," Romm told the House Science Committee in March in written testimony.

To combat the warming threat, funding agencies should place a near-term priority on promoting energy efficiency, research on renewables, and development of hybrid cars, critics say. After all, many researchers point out, as long as hydrogen for fuel cell cars is provided from fossil fuels, much the same environmental benefits can be gained by adopting hybrid gasoline-electric and advanced diesel engines. As MIT chemist and former DOE director of energy research John Deutch and colleagues point out on page 974, hybrid electric vehicles—a technology already on the market—would improve energy efficiency and reduce greenhouse gas emissions almost as well as fuel cell vehicles that generate hydrogen from an onboard gasoline reformer, an approach that obviates the need for building a separate hydrogen infrastructure.

Near-term help may also come from capturing CO_2 emissions from power and industrial plants and storing them underground, a process known as carbon sequestration (see p. 962). Research teams from around the world are currently testing a variety of schemes for doing that. But the process remains significantly more expensive than current energy. "Until an economical solution to the sequestration problem is found, net reductions in overall CO_2 emissions can only come through advances in energy efficiency and renewable energy," the APS report concludes.

In response to the litany of concerns over making the transition to a hydrogen economy, JoAnn Milliken, who heads hydrogen-storage research for DOE, points out that DOE and other funding agencies aren't promoting hydrogen to the exclusion of other energy research. Renewable energy, carbon sequestration, and even fusion energy all remain in the research mix. Criticism that too much is being spent on demonstration projects is equally misguided, she says, noting that such projects make up only 13% of DOE's hydrogen budget, compared with 85% for basic and applied research. Both are necessary, she says: "We've been doing basic research on hydrogen for a long time. We can't just do one or the other." Finally, she points out, funding agencies have no illusions about the challenge in launching the hydrogen economy. "We never said this is going to be easy," Milliken says. The inescapable truth is that "we need a substitute for gasoline. Gas hybrids are going to improve fuel economy. But they can't solve the problem."

Yet, if that's the case, many energy experts argue, governments should be spending far more money to lower the technical and economic barriers to all types of alternative energy—hydrogen included—and bring it to reality sooner. "Energy is the single most important problem facing humanity today," says Richard Smalley of Rice University in Houston, Texas, a 1996 Nobel laureate in chemistry who has been campaigning for increased energy sciences funding for the last 2 years. Among Smalley's proposals: a 5-cent-per-gallon tax on gasoline in the United States to fund $10 billion annually in basic energy sciences research. Because of the combination of climate change and the soon-to-be-peaking production in fossil fuels, Smalley says, "it really ought to be the top project in worldwide science right now."

Although not all researchers are willing to wade into the political minefield of backing a gasoline tax, few disagree with his stand. "I think he's right," Dresselhaus says of the need to boost the priority of basic energy sciences research. With respect to the money needed to take a realistic stab at making an alternative energy economy a reality, Dresselhaus says: "Most researchers think there isn't enough money being spent. I think the investment is pretty small compared to the job that has to be done." Even though it sounds like a no-brainer, the hydrogen economy will take abundant gray matter and greenbacks to bring it to fruition.

–ROBERT F. SERVICE

NEWS

The Carbon Conundrum

En route to hydrogen, the world will have to burn huge amounts of fossil fuels—and find ways to deal with their climate-changing byproducts

Even if the hydrogen economy were technically and economically feasible today, weaning the world off carbon-based fossil fuels would still take decades. During that time, carbon combustion will continue to pour greenhouse gases into the atmosphere—unless scientists find a way to reroute them. Governments and energy companies around the globe have launched numerous large-scale research and demonstration projects to capture and store, or sequester, unwanted carbon dioxide (see table). Although final results are years off, so far the tests appear heartening. "It seems to look more and more promising all the time," says Sally Benson, a hydrogeologist at Lawrence Berkeley National Laboratory in California. "For the first time, I think the technical feasibility has been established."

Burial at sea. A million tons a year of CO_2 from the Sleipner natural-gas field in the North Sea are reinjected underground.

Last hope?

Fossil fuels account for most of the 6.5 billion tons (gigatons) of carbon—the amount present in 25 gigatons of CO_2—that people around the world vent into the atmosphere every year. And as the amount of the greenhouse gas increases, so does the likelihood of triggering a debilitating change in Earth's climate. Industrialization has already raised atmospheric CO_2 levels from 280 to 370 parts per million, which is likely responsible for a large part of the 0.6°C rise in the average global surface temperature over the past century. As populations explode and economies surge, global energy use is expected to rise by 70% by 2020, according to a report last year from the European Commission, much of it to be met by fossil fuels. If projections of future fossil fuel use are correct and nothing is done to change matters, CO_2 emissions will increase by 50% by 2020.

To limit the amount of CO_2 pumped into the air, many scientists have argued for capturing a sizable fraction of that CO_2 from electric plants, chemical factories, and the like and piping it deep underground. In June, Ronald Oxburgh, Shell's chief in the United Kingdom, called sequestration essentially the last best hope to combat climate change. "If we don't have sequestration, then I see very little hope for the world," Oxburgh told the British newspaper *The Guardian*.

Although no one has adopted the strategy on a large scale, oil companies have been piping CO_2 underground for decades to extract more oil from wells by reducing the viscosity of underground oil. Because they weren't trying to maximize CO_2 storage, companies rarely tracked whether the CO_2 remained underground or caused unwanted side effects.

That began to change in the early 1990s, when researchers began to consider sequestering CO_2 to keep it out of the atmosphere. The options for doing so are limited, says Robert Kane, who heads carbon-sequestration programs at the U.S. Department of Energy in Washington, D.C. You can grow plants that consume CO_2 to fuel their growth, or pipe the gas to the deep ocean or underground. But planted vegetation can burn or be harvested, ultimately returning the CO_2 back into the atmosphere. And placing vast amounts of CO_2 into the ocean creates an acidic plume, which can wreak havoc on deep-water ecosystems (*Science*, 3 August 2001, p. 790). As a result, Kane and others say, much recent research has focused on storing the CO_2 underground in depleted oil and gas reservoirs, coal seams that are too deep to mine, and underground pockets of saltwater called saline aquifers.

"Initially, it sounded like a wild idea," Benson says, in part because the volume of gas that would have to be stored is enormous. For example, storing just 1 gigaton of CO_2—about 4% of what we vent annually worldwide—would require moving 4.8 million cubic meters of gas a day, equivalent to about one-third the volume of all the oil shipped daily around the globe. But early studies suggest that there is enough underground capacity to store hundreds of years' worth of CO_2 injection, and that potential underground storage sites exist worldwide. According to Benson, studies in the mid-1990s pegged the underground storage capacity between 1000 and 10,000 gigatons of CO_2. More detailed recent analyses are beginning to converge around the middle of that range, Benson says. But even the low end is comfortably higher than the 25 gigatons of CO_2 humans produce each year, she notes.

To test the technical feasibility, researchers have recently begun teaming up with oil and gas companies to study their CO_2 piping projects. One of the first, and the biggest, is the Weyburn project in Saskatchewan, Canada. The site is home to an oil field discovered in 1954. Since then, about one-quarter of the reservoir's oil has been removed, producing 1.4 billion barrels. In 1999, the Calgary-based oil company EnCana launched a $1.5 billion, 30-year effort to pipe 20 million metric tons of CO_2 into the reservoir after geologists estimated that it would increase the field's yield by another third. For its CO_2, EnCana teamed up with the Dakota Gasification Co., which operates a plant in Beulah, North Dakota, that converts coal into a hydrogen-rich gas used in industry and that emits CO_2 as a byproduct. EnCana built a 320-km pipeline to carry pressurized CO_2 to Weyburn, where it's injected underground.

In September 2000, EnCana began injecting an estimated 5000 metric tons of CO_2 a

CREDIT: STATOIL

day 1500 meters beneath the surface. The technology essentially just uses compressors to force compressed CO_2 down a long pipe drilled into the underground reservoir. To date, nearly 3.5 million metric tons of CO_2 have been locked away in the Weyburn reservoir.

When the project began, the United States was still party to the Kyoto Protocol, the international treaty designed to reduce greenhouse gas emissions. So the United States, Canada, the European Union, and others funded $28 million worth of modeling, monitoring, and geologic studies to track the fate of Weyburn's underground CO_2.

For the first phase of that study, which ended in May, 80 researchers including geologists and soil scientists monitored the site for 4 years. "The short answer is it's working," says geologist and Weyburn team member Ben Rostron of the University of Alberta in Edmonton: "We've got no evidence of significant amounts of injected CO_2 coming out at the surface." That was what they expected, Rostron says: Wells are sealed and capped, and four layers of rock thought to be impermeable to CO_2 lie between the oil reservoir and the surface.

A similar early-stage success story is under way in the North Sea off the coast of Norway. Statoil, Norway's largest oil company, launched a sequestration pilot project from an oil rig there in 1996 to avoid a $55-a-ton CO_2 tax that the Norwegian government levies on energy producers. The rig taps a natural gas field known as Sleipner, which also contains large amounts of CO_2. Normally, gas producers separate the CO_2 from the natural gas before feeding the latter into a pipeline or liquefying it for transport. The CO_2 is typically vented into the air. But for the past 8 years, Statoil has been injecting about 1 million tons of CO_2 a year back into a layer of porous sandstone, which lies between 550 and 1500 meters beneath the ocean floor. Sequestering the gas costs about $15 per ton of CO_2 but saves the company $40 million a year in tax.

Researchers have monitored the fate of the CO_2 with the help of seismic imaging and other tools. So far, says Stanford University petroleum engineer Franklin Orr, everything suggests that the CO_2 is staying put. Fueled by these early successes, other projects are gearing up as well. "One can't help but be struck by the dynamism in this community right now," says Princeton University sequestration expert Robert Socolow. "There is a great deal going on."

Despite the upbeat early reviews, most researchers and observers are cautious about the prospects for large-scale sequestration. "Like every environmental issue, there are certain things that happen when the quantity increases," Socolow says. "We have enough history of getting this [type of thing] wrong that everyone is wary."

Safety tops the concerns. Although CO_2 is nontoxic (it constitutes the bubbles in mineral water and beer), it can be dangerous. If it percolates into a freshwater aquifer, it can acidify the water, potentially leaching lead, arsenic, or other dangerous trace elements into the mix. If the gas rises to the subsurface, it can affect soil chemistry. And if it should escape above ground in a windless depression, the heavier-than-air gas could collect and suffocate animals or people. Although such a disaster hasn't happened yet with sequestered CO_2, the threat became tragically clear in 1986, when an estimated 80 million cubic meters of CO_2 erupted from the Lake Nyos crater in Cameroon, killing 1800 people.

Money is another issue. Howard Herzog, an economist at the Massachusetts Institute of Technology in Cambridge and an expert on the cost of sequestration, estimates that large-scale carbon sequestration would add 2 to 3 cents per kilowatt-hour to the cost of electricity delivered to the consumer—about one-third the average cost of residential electricity in the United States. (A kilowatt-hour of electricity can power 10 100-watt light bulbs for an hour.) Says Orr: "The costs are high enough that this won't happen on a big scale without an incentive structure" such as Norway's carbon tax or an emissions-trading program akin to that used with sulfur dioxide, a component of acid rain.

But although sequestration may not be cheap, Herzog says, "it's affordable." Generating electricity with coal and storing the carbon underground still costs only about 14% as much as solar-powered electricity. And unlike most renewable energy, companies can adopt it more easily on a large scale and can retrofit existing power plants and chemical plants. That's particularly important for dealing with the vast amounts of coal that are likely to be burned as countries such as China and India modernize their economies. "Coal is not going to go away," Herzog says. "People need energy, and you can't make energy transitions easily." Sequestration, he adds, "gives us time to develop 22nd century energy sources." That could give researchers a window in which to develop and install the technologies needed to power the hydrogen economy.

–ROBERT F. SERVICE

Some CO_2 Sequestration Projects

Project	Location	Tons of CO_2 to be injected	Source	Status
Sleipner	North Sea	20 million	Gas field	Ongoing
Weyburn	Canada	20 million	Oil field	Completed phase 1
In Salah	Algeria	18 million	Gas field	Starts 2004
Gorgon	Australia	125 million	Saline aquifer	In preparation
Frio	U.S.	3000	Saline aquifer	Pilot phase
RECOPOL	Poland	3000	Coal seams	Ongoing

Choosing a CO_2 Separation Technology

If governments move to deep-six carbon dioxide, much of the effort is likely to target emissions from coal-fired power plants. Industrial companies have used detergent-like chemicals and solvents for decades to "scrub" CO_2 from flue gases, a technique that can be applied to existing power plants. The downside is that the technique is energy intensive and reduces a coal plant's efficiency by as much as 14%. Another option is to burn coal with pure oxygen, which produces only CO_2 and water vapor as exhaust gases. The water vapor can then be condensed, leaving just the CO_2. But this technology too consumes a great deal of energy to generate the pure oxygen in the first place and reduces a coal plant's overall efficiency by about 11%. A third approach extracts CO_2 from coal before combustion. This technique is expected to be cheaper and more efficient, but it requires building plants based on a newer technology, known as Integrated Gasification Combined Cycle. But it will take a carbon tax or some other incentive to drive utility companies away from proven electricity-generating technology. **–R.F.S.**

Dark victory. Coal can be made cleaner, for a price.

CREDIT: (BOTTOM) ROYALTY-FREE/CORBIS

NEWS

Fire and ICE: Revving Up for H_2

The first hydrogen-powered cars will likely burn the stuff in good old internal combustion engines. But can they drive the construction of hydrogen infrastructure?

In the day we sweat it out in the streets of
a runaway American dream.
At night we ride through mansions of
glory in suicide machines,
Sprung from cages out on highway 9,
Chrome wheeled, fuel injected
and steppin' out over the line …

Fear not, sports car aficionados and Bruce Springsteen fans: Even if the hydrogen economy takes off, it may be decades before zero-emission fuel cells replace your beloved piston-pumping, fuel-burning, song-inspiring internal combustion engine. In the meantime, however, instead of filling your tank with gasoline, you may be pumping hydrogen.

A handful of automakers are developing internal combustion engines that run on hydrogen, which burns more readily than gasoline and produces almost no pollutants. If manufacturers can get enough of them on the road in the next few years, hydrogen internal combustion engine (or H2 ICE) vehicles might spur the construction of a larger infrastructure for producing and distributing hydrogen—the very same infrastructure that fuel cell vehicles will require.

If all goes as hoped, H2 ICE vehicles could solve the chicken-or-the-egg problem of which comes first, the fuel cell cars or the hydrogen stations to fuel them, says Robert Natkin, a mechanical engineer at Ford Motor Co. in Dearborn, Michigan. "The prime reason for doing this is to get the hydrogen economy under way as quickly as possible," Natkin says. In fact, some experts say that in the race to economic and technological viability, the more cumbersome, less powerful fuel cell may never catch up to the lighter, peppier, and cheaper H2 ICE. "If the hydrogen ICEs work the way we think they can, you may never see fuel cells" powering cars, says Stephen Ciatti, a mechanical engineer at Argonne National Laboratory in Illinois.

BMW, Ford, and Mazda expect to start producing H2 ICE vehicles for government and commercial fleets within a few years. But to create demand for hydrogen, those cars and trucks will have to secure a niche in the broader consumer market, and that won't be a drive in the countryside. The carmakers have taken different tacks to keeping hydrogen engines running smoothly and storing enough hydrogen onboard a vehicle to allow it to wander far from a fueling station, and it remains to be seen which approach will win out. And, of course, H2 ICE vehicles will require fueling stations, and most experts agree that the public will have to help pay for the first ones.

Most important, automakers will have to answer a question that doesn't lend itself to simple, rational analysis: At a time when gasoline engines run far cleaner than they once did and sales of gas-guzzling sport utility vehicles continue to grow in spite of rising oil prices, what will it take to put the average driver behind the wheel of an exotic hydrogen-burning car?

Motoring. Hydrogen engines, such as the one that powers Ford's Model U concept car, may provide the technological steppingstone to fuel-cell vehicles.

Running lean and green

An internal combustion engine draws its power from a symphony of tiny explosions in four beats. Within an engine, pistons slide up and down within snug-fitting cylinders. First, a piston pushes up into its cylinder to compress a mixture of air and fuel. When the piston nears the top of its trajectory, the sparkplug ignites the vapors. Next, the explosion pushes the piston back down, turning the engine's crankshaft and, ultimately, the wheels of the car. Then, propelled by inertia and the other pistons, the piston pushes up again and forces the exhaust from the explosion out valves in the top of the cylinder. Finally, the piston descends again, drawing a fresh breath of the air-fuel mixture into the cylinder through a different set of valves and beginning the four-stroke cycle anew.

A well-tuned gasoline engine mixes fuel and air in just the right proportions to ensure that the explosion consumes essentially every molecule of fuel and every molecule of oxygen—a condition known as "running at stoichiometry." Of course, burning gasoline produces carbon monoxide, carbon dioxide, and hydrocarbons. And when running at stoichiometry, the combustion is hot enough to burn some of the nitrogen in the air, creating oxides of nitrogen (NO_x), which seed the brown clouds of smog that hang over Los Angeles and other urban areas.

In contrast, hydrogen coughs up almost no pollutants. Burning hydrogen produces no carbon dioxide, the most prevalent heat-trapping greenhouse gas, or other carbon compounds. And unlike gasoline, hydrogen burns even when the air-fuel mixture contains far less hydrogen than is needed to consume all the oxygen—a condition known as "running lean." Reduce the hydrogen-air mixture to roughly half the stoichiometric ratio, and the temperature of combustion falls low enough to extinguish more than 90% of NO_x production. Try that with a gasoline engine and it will run poorly, if at all.

But before they can take a victory lap, engineers working on H2 ICEs must solve some problems with engine performance. Hydrogen packs more energy per kilogram than gasoline, but it's also the least dense gas in nature, which means it takes up a lot of room in an engine's cylinders, says Christopher White, a mechanical engineer at Sandia National Laboratories in Livermore, California. "That costs you power because there's less oxygen to consume," he says. At the same time, it takes so little energy to ig-

nite hydrogen that the hydrogen-air mixture tends to go off as soon as it gets close to something hot, like a sparkplug. Such "preignition" can make an engine "knock" or even backfire like an old Model T.

Power play

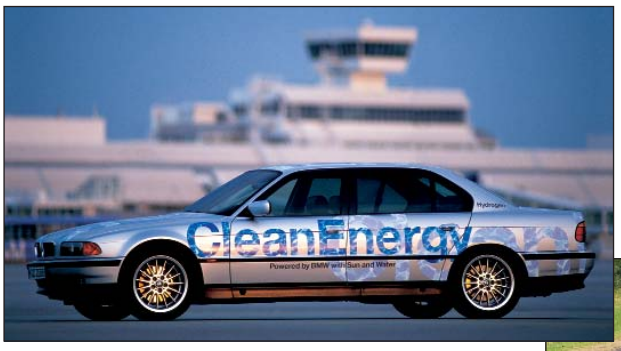
To surmount such problems, BMW, Ford, and Mazda are taking different tacks. Ford engineers use a mechanically driven pump called a supercharger to force more air and fuel into the combustion chamber, increasing the energy of each detonation. "We basically stuff another one-and-a-half times more air, plus an appropriate amount of fuel, into the cylinders," says Ford's Natkin. Keeping the hydrogen-air ratio very lean—less than 40% of the stoichiometric ratio—prevents preignition and backfire, he says. A hydrogen-powered Focus compact car can travel about 240 kilometers before refueling its hydrogen tanks, which are pressurized to 350 times atmospheric pressure. And with an electric hybrid system and tanks pressurized to 700 atmospheres, or 70 MPa, Ford's Model U concept car can range twice as far.

Mazda's H2 ICE prototype also carries gaseous hydrogen, but it burns it in a rotary engine driven by two triangular rotors. To overcome hydrogen's propensity to displace air, Mazda engineers force the hydrogen into the combustion chamber only after the chamber has filled with air and the intake valves have closed. As well as boosting power, such "direct injection" eliminates backfiring by separating the hydrogen and oxygen until just before they're supposed to detonate, explains Masanori Misumi, an engineer at Mazda Motor Corp. in Hiroshima, Japan. Mazda's hydrogen engine will also run on gasoline.

When BMW's H2 ICE needs maximum power, it pours on the hydrogen to run at stoichiometry. Otherwise, it runs lean. A hydrogen-powered Beemer also carries denser liquid hydrogen, boiling away at –253ºC inside a heavily insulated tank, which greatly increases the distance a car can travel between refueling stops. In future engines, the chilly gas might cool the air-fuel mixture, making it denser and more potent than a warm mixture. The cold hydrogen gas might also cool engine parts, preventing backfire and preignition. BMW's H2 ICE can run on gasoline as well.

Unlike Ford and Mazda, BMW has no immediate plans to pursue fuel cell technology alongside its H2 ICEs. A fuel cell can wring more useful energy from a kilogram of hydrogen, but it cannot provide the wheel-spinning power that an internal combustion engine can, says Andreas Klugescheid, a spokesperson for the BMW Group in Munich. "Our customers don't buy a car just to get from A to B, but to have fun in between," he says. "At BMW we're pretty sure that the hydrogen internal combustion engine is the way to satisfy them."

The first production H2 ICE vehicles will likely roll off the assembly line within 5 years, although the automakers won't say precisely when. "We would anticipate a lot more hydrogen internal combustion engine vehicles on the road sooner rather than later as we continue to develop fuel cell vehicles," says Michael Vaughn, a spokesperson for Ford in Dearborn. Focusing on the market for luxury performance cars, BMW plans to produce some hydrogen-powered vehicles in the current several-year model cycle of its flagship 7 Series cars. Automakers will introduce the cars into commercial and government fleets, taking advantage of the centralized fueling facilities to carefully monitor their performance.



What a gas! H2 ICE vehicles, such as BMW's and Mazda's prototypes, promise performance as well as almost zero emissions.

Supplying demand

In the long run, most experts agree, the hydrogen fuel cell holds the most promise for powering clean, ultraefficient cars. If they improve as hoped, fuel cells might usefully extract two-thirds of the chemical energy contained in a kilogram of hydrogen. In contrast, even with help from an electric hybrid system, an H2 ICE probably can extract less than half. (A gasoline engine makes use of about 25% of the energy in its fuel, the rest going primarily to heat.) And whereas an internal combustion engine will always produce some tiny amount of pollution, a fuel cell promises true zero emissions. But H2 ICE vehicles enjoy one advantage that could bring them to market quickly and help increase the demand for hydrogen filling stations, says James Francfort, who manages the Department of Energy's Advanced Vehicle Testing Activity at the Idaho National Engineering and Environmental Laboratory in Idaho Falls. "The car guys know how to build engines," Francfort says. "This looks like something that could be done now."

Developers are hoping that H2 ICE vehicles will attract enough attention in fleets that a few intrepid individuals will go out of their way to buy them, even if they have to fuel them at fleet depots and the odd publicly funded fueling station. If enough people follow the trendsetters, eventually the demand for hydrogen refueling stations will increase to the point at which they become profitable to build and operate—or so the scenario goes.

All agree that if H2 ICEs are to make it out of fleets and into car dealers' showrooms, they'll need a push from the public sector. They're already getting it in California, which gives manufacturers environmental credits for bringing H2 ICE vehicles to market. And in April, California Governor Arnold Schwarzenegger announced a plan to line California's interstate highways with up to 200 hydrogen stations by 2010, just in time to kick-start the market for H2 ICEs.

Ultimately, the fate of H2 ICE vehicles lies with consumers, who have previously turned a cold shoulder to alternative technologies such as cars powered by electricity, methanol, and compressed natural gas. With near-zero emissions and an edge in power over fuel cells, the H2 ICE might catch on with car enthusiasts yearning to go green, a demographic that has few choices in today's market, says BMW's Klugescheid. If the H2 ICE can help enough gearheads discover their inner tree-hugger, the technology might just take off. "There are enough people who are deeply in love with performance cars but also have an environmental conscience," Klugescheid says.

Developers hope that the H2 ICE vehicles possess just the right mixture of environmental friendliness, futuristic technology, and good old-fashioned horsepower to capture the imagination of the car-buying public. A few short years should tell if they do. In the meantime, it wouldn't hurt if Bruce Springsteen wrote a song about them, too.

–ADRIAN CHO

CREDITS: (LEFT TO RIGHT) BMW AG; MAZDA MOTOR CORP.

NEWS

Will the Future Dawn in the North?

With geothermal and hydroelectric sources supplying almost all of its heat and electricity, Iceland is well on the way to energy self-suffiency. Now it is betting that hydrogen-fueled transportation will supply the last big piece of the puzzle

REYKJAVIK—As commuters in this coastal city board the route 2 bus, some get an unexpected chemistry lesson. The doors of the bus are emblazoned with a brightly painted diagram of a water molecule, H_2O. When they swing open, oxygen goes right and hydrogen left, splitting the molecule in two. The bus's sponsors hope that soon all of Iceland's nearly 300,000 residents not only will know this chemical reaction but will be relying on it to get around: By 2050, they say, Iceland should run on a completely hydrogen-based energy economy.

The hydrogen-powered fuel cell buses—three ply the route regularly—are the first step toward weaning Iceland off imported fossil fuels, says Thorsteinn Sigfusson, a physicist at the University of Iceland in Reykjavik and founding chair of Icelandic New Energy, a company launched to test and develop hydrogen-based transport and energy systems.

Although other European countries are fielding similar pilot projects, this volcanic island nation is uniquely poised to tap hydrogen's potential. Iceland already uses water—either hot water from geothermal sources or falling water from hydroelectric dams—to provide hot showers, heat its homes, and light its streets and buildings through the long, dark winter just south of the Arctic Circle. "This is a sustainable Texas," says Sigfusson, referring to the plentiful energy welling from the ground. But although the country has the capacity to produce much more energy than it currently needs, it still imports 30% of its energy as oil to power cars and ships. Converting those vehicles to hydrogen fuel could make the country self-sufficient in energy.

Iceland started taking the idea of a hydrogen economy seriously more than a decade ago, after the 1990 Kyoto Protocol required that it cut its nonindustrial CO_2 emissions. Having already converted more than 95% of its heat and electricity generation to hydroelectric and geothermal energy, which emit almost no CO_2, the country focused on the transportation sector. Spurred by an expert commission that recommended hydrogen fuels as the most promising way to convert its renewable energy into a fuel to power cars, the government formed Icelandic New Energy in 1997. Today 51% of the shares are owned by Icelandic sources, including power companies, the University of Iceland, and the government itself. The rest are held by international corporations Norsk Hydro, DaimlerChrysler, and Shell Hydrogen. With $4 million from the European Union and $5 million from its investors, the company bought three hydrogen-powered buses for the city bus fleet and built a fueling station to keep them running.

All aboard. Fuel cell buses and a planned car fleet are the latest entries in Iceland's marathon push for total energy independence.

The fueling station is part of the country's busiest Shell gasoline station, clearly visible from the main road out of Reykjavik. It boldly proclaims to all passersby in English that hydrogen is "the ultimate fuel" and "We're making the fuel of the future." The hydrogen is produced overnight using electricity from the city's power grid to split water into hydrogen and oxygen. The hydrogen is then stored as a compressed gas. It takes just over 6 minutes to fill a bus with enough fuel to run a full day's journey. The buses are built from standard bodies outfitted with rooftop hydrogen tanks that make them slightly taller than their diesel cousins. "The first cars were horse carriages retrofitted with engines. We're seeing the same process here," Sigfusson says. Piped to the fuel cells, the hydrogen combines with oxygen from the air and produces electricity, which drives a motor behind the rear wheels.

As the buses tour the city, their exhaust pipes emit trails of steam strikingly reminiscent of those that waft from the hot springs in the mountains outside the city. The exhaust is more visible than that from other cars, Sigfusson says, because the fuel cell runs at about 70°C and so the steam is close to the saturation point. But it is almost pure water, he says, so clean "you can drink it." And because the buses are electric, they are significantly quieter than diesel buses.

The pilot project has not been trouble-free. On a recent Friday, for example, two of the three buses were out of service for repairs. The buses are serviced in a garage equipped with special vents that remove any highly explosive hydrogen that might escape while a bus is being repaired. On cold winter nights the vehicles must be kept warm in specially designed bays, lest the water vapor left in the system freeze and damage the fuel cells. "They need to be kept like stallions in their stalls," says Sigfusson, who notes that newer generations of fuel cells are drier and may not need such coddling.

But despite the hiccups, Sigfusson says the project so far has been encouraging. In 9 months, the buses have driven a total of 40,000 kilometers, while surveys show that public support for a hydrogen economy has remained at a surprising 93%. The next step is a test fleet of passenger cars, Sigfusson says. Icelandic New Energy is negotiating to buy more than a dozen hydrogen-powered cars for corporate or government employees, he says.

Economic leaders are also optimistic. "I am not a believer that we will have a hydrogen economy tomorrow," says Friðrik Sophusson, a former finance minister and now managing director of the government-owned National Power Co., a shareholder in Icelandic New Energy. But he believes the investment will pay long-term dividends—not least to his company, which will supply electricity needed to produce the gas. "In 20 years, I believe we will have vehicles running on hydrogen efficiently generated from renewable sources," Sophusson says. "We are going to produce hydrogen in a clean way, and if the project takes off, we will be in business."

Although Iceland may harbor the most

CREDIT: COURTESY OF ICELANDIC NEW ENERGY

ambitious vision of a fossil fuel–free future, other countries without its natural advantages in renewable energy are also experimenting with hydrogen-based technologies. Sigfusson thinks the gas's biggest potential could lie in developing countries that have not yet committed themselves to fossil fuels (see sidebar), but industrialized nations are also pushing hard. In a project partly funded by the European Union (E.U.), nine other European cities now have small fleets of buses—similar to the ones in Reykjavik—plying regular routes. The E.U. imports 50% of its oil, and that figure is expected to rise to 70% over the next 20 to 30 years. In January, European Commission President Romano Prodi pledged to create "a fully integrated hydrogen economy, based on renewable energy sources, by the middle of the century."

Realizing that bold ambition is now the job of the Hydrogen and Fuel Cell Technology Platform, an E.U. body. Its advisory panel of 35 prominent industry, research, and civic leaders will coordinate efforts in academia and industry at both the national and European levels and will draw up a research plan and deployment strategy. Planned demonstration projects include a fossil fuel power plant that will produce electricity and hydrogen on an industrial scale while separating and storing the CO_2 it produces and a "hydrogen village" where new technologies and hydrogen infrastructure can be tested. In all, the E.U.'s Framework research program intends to spend $300 million on hydrogen and fuel cell research during the 5-year period from 2002 to 2006. Political and public interest in a hydrogen economy "is like a snowball growing and growing," says Joaquín Martín Bernejo, the E.U. official responsible for research into hydrogen production and distribution.

Fill 'er up. Reykjavik's lone hydrogen station, which manufactures the fuel on site by splitting water molecules, can fill a bus with pressurized gas in 6 minutes.

As Iceland (not an E.U. member) treads that same path on a more modest scale, its biggest hurdle remains the conversion of its economically vital shipping fleet, which uses half of the country's imported oil. Boats pose tougher technical problems than city buses do. Whereas a bus can run its daily route on only 40 kilograms of hydrogen, Sigfusson says, a small trawler with a 500-kilowatt engine must carry a ton of the gas to spend 4 to 5 days at sea. One way to store enough fuel, he says, might be to sequester the gas in hydrogen-absorbing compounds called metal hydrides. The compound could even serve as ballast for the boat instead of the standard concrete keel ballast.

Although Iceland's leaders are eager for the hydrogen economy to take off, Sigfusson acknowledges that it will have to appeal directly to Iceland's drivers and fishers. Generous tax breaks to importers of hydrogen vehicles will help, he says, if hydrogen can match the price of heavily taxed fossil fuels: "People will be willing to pay a little more [for a hydrogen vehicle], but they're not willing to pay a lot more. The market has to force down the price of an installed kilowatt." According to Sigfusson, that is already happening, especially as research investments are ramped up: "There has been a paradigm shift. We had had decades of coffee-room discussions that never led anywhere." Now the buses with their chemistry lesson, he says, are pointing the way to the future.

–GRETCHEN VOGEL

With reporting by Daniel Clery.

Can the Developing World Skip Petroleum?

If technologies for hydrogen fuel take off, one of the biggest winners could be the developing world. Just as cell phones in poor countries have made land lines obsolete before they were installed, hydrogen from renewable sources—in an ideal world—could enable developing countries to leap over the developed world to energy independence. "The opportunity is there for them to become leaders in this area," says Thorsteinn Sigfusson of the University of Iceland, one of the leaders of the International Partnership for a Hydrogen Economy (IPHE), a cooperative effort of 15 countries, including the United States, Iceland, India, China, and Brazil, founded last year to advance hydrogen research and technology development.

With their growing influence in global manufacturing, their technical expertise, and their low labor costs, Sigfusson says, countries such as China and India could play extremely important roles in developing more efficient solar or biotech sources of hydrogen—as well as vehicles and power systems that use the fuel. "They have the opportunity to take a leap into the hydrogen economy without all the troubles of going through combustion and liquid fuel," he says. The impact would be huge. The IPHE members already encompass 85% of the world's population, he notes.

The current steps are small. For example, a joint U.S.-Indian project is working to build a hydrogen-powered three-wheel test vehicle. The minicar, designed for crowded urban streets, needs only one-tenth as much storage space as a standard passenger car. India's largest auto manufacturer, Mahindra and Mahindra, has shipped two of its popular gasoline-powered three-wheelers (currently a huge source of urban air pollution), to the Michigan-based company Energy Conversion Devices (ECD). Engineers at ECD are working to convert the engine to run on hydrogen stored in a metal hydride. One model will return to India for testing, and one will remain in the United States. The small project "is just the beginning," says a U.S. Department of Energy official. "But the point of bringing in these countries is that they are huge energy consumers. They simply have to be part of the partnership, especially as we start to use the technologies."

Different future. Countries not yet committed to fossil fuels might go straight to hydrogen.

Ideally, the developing world will be able to harness the solar energy plentiful in the tropics to power hydrogen systems, Sigfusson says. "The most important renewable will be the sun," he says. "Mankind started as a solar civilization. We spent 200 years flirting with fossil fuels, but I believe we'll soon go back to being a solar civilization." **–G.V.**

CREDITS: (TOP TO BOTTOM) G. VOGEL; SAURABH DAS/AP PHOTO

REVIEW

Stabilization Wedges: Solving the Climate Problem for the Next 50 Years with Current Technologies

S. Pacala[1]* and R. Socolow[2]*

Humanity already possesses the fundamental scientific, technical, and industrial know-how to solve the carbon and climate problem for the next half-century. A portfolio of technologies now exists to meet the world's energy needs over the next 50 years and limit atmospheric CO_2 to a trajectory that avoids a doubling of the preindustrial concentration. Every element in this portfolio has passed beyond the laboratory bench and demonstration project; many are already implemented somewhere at full industrial scale. Although no element is a credible candidate for doing the entire job (or even half the job) by itself, the portfolio as a whole is large enough that not every element has to be used.

The debate in the current literature about stabilizing atmospheric CO_2 at less than a doubling of the preindustrial concentration has led to needless confusion about current options for mitigation. On one side, the Intergovernmental Panel on Climate Change (IPCC) has claimed that "technologies that exist in operation or pilot stage today" are sufficient to follow a less-than-doubling trajectory "over the next hundred years or more" [(*1*), p. 8]. On the other side, a recent review in *Science* asserts that the IPCC claim demonstrates "misperceptions of technological readiness" and calls for "revolutionary changes" in mitigation technology, such as fusion, space-based solar electricity, and artificial photosynthesis (*2*). We agree that fundamental research is vital to develop the revolutionary mitigation strategies needed in the second half of this century and beyond. But it is important not to become beguiled by the possibility of revolutionary technology. Humanity can solve the carbon and climate problem in the first half of this century simply by scaling up what we already know how to do.

What Do We Mean by "Solving the Carbon and Climate Problem for the Next Half-Century"?

Proposals to limit atmospheric CO_2 to a concentration that would prevent most damaging climate change have focused on a goal of 500 ± 50 parts per million (ppm), or less than double the preindustrial concentration of 280 ppm (*3*–*7*). The current concentration is ~375 ppm. The CO_2 emissions reductions necessary to achieve any such target depend on the emissions judged likely to occur in the absence of a focus on carbon [called a business-as-usual (BAU) trajectory], the quantitative details of the stabilization target, and the future behavior of natural sinks for atmospheric CO_2 (i.e., the oceans and terrestrial biosphere). We focus exclusively on CO_2, because it is the dominant anthropogenic greenhouse gas; industrial-scale mitigation options also exist for subordinate gases, such as methane and N_2O.

Very roughly, stabilization at 500 ppm requires that emissions be held near the present level of 7 billion tons of carbon per year (GtC/year) for the next 50 years, even though they are currently on course to more than double (Fig. 1A). The next 50 years is a sensible horizon from several perspectives. It is the length of a career, the lifetime of a power plant, and an interval for which the technology is close enough to envision. The calculations behind Fig. 1A are explained in Section 1 of the supporting online material (SOM) text. The BAU and stabilization emissions in Fig. 1A are near the center of the cloud of variation in the large published literature (*8*).

The Stabilization Triangle

We idealize the 50-year emissions reductions as a perfect triangle in Fig. 1B. Stabilization is represented by a "flat" trajectory of fossil fuel emissions at 7 GtC/year, and BAU is represented by a straight-line "ramp" trajectory rising to 14 GtC/year in 2054. The "stabilization triangle," located between the flat trajectory and BAU, removes exactly one-third of BAU emissions.

To keep the focus on technologies that have the potential to produce a material difference by 2054, we divide the stabilization triangle into seven equal "wedges." A wedge represents an activity that reduces emissions to the atmosphere that starts at zero today and increases linearly until it accounts for 1 GtC/year of reduced carbon emissions in 50 years. It thus represents a cumulative total of 25 GtC of reduced emissions over 50 years. In this paper, to "solve the carbon and climate problem over the next half-century" means to deploy the technologies and/or lifestyle changes necessary to fill all seven wedges of the stabilization triangle.

Stabilization at any level requires that net emissions do not simply remain constant, but eventually drop to zero. For example, in one simple model (*9*) that begins with the stabilization triangle but looks beyond 2054, 500-ppm stabilization is achieved by 50 years of flat emissions, followed by a linear decline of about two-thirds in the following 50 years, and a very slow decline thereafter that matches the declining ocean sink. To develop the revolutionary technologies required for such large emissions reductions in the second half of the century, enhanced research and development would have to begin immediately.

Policies designed to stabilize at 500 ppm would inevitably be renegotiated periodically to take into account the results of research and development, experience with specific wedges, and revised estimates of the size of the stabilization triangle. But not filling the stabilization triangle will put 500-ppm stabilization out of reach. In that same simple model (*9*), 50 years of BAU emissions followed by 50 years of a flat trajectory at 14 GtC/year leads to more than a tripling of the preindustrial concentration.

It is important to understand that each of the seven wedges represents an effort beyond what would occur under BAU. Our BAU simply continues the 1.5% annual carbon emissions growth of the past 30 years. This historic trend in emissions has been accompanied by 2% growth in primary energy consumption and 3% growth in gross world product (GWP) (Section 1 of SOM text). If carbon emissions were to grow 2% per year, then ~10 wedges would be needed instead of 7, and if carbon emissions were to grow at 3% per year, then ~18 wedges would be required (Section 1 of SOM text). Thus, a continuation of the historical rate of decarbonization of the fuel mix prevents the need for three additional wedges, and ongoing improvements in energy efficiency prevent the need for eight additional wedges. Most readers will reject at least one of the wedges listed here, believing that the corresponding deployment is certain to occur in BAU, but readers will disagree about which to reject on such grounds. On the other hand, our list of mitigation options is not exhaustive.

[1]Department of Ecology and Evolutionary Biology, [2]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA.

*To whom correspondence should be addressed. E-mail: pacala@princeton.edu (S.P.); socolow@princeton.edu (R.S.)

What Current Options Could Be Scaled Up to Produce at Least One Wedge?

Wedges can be achieved from energy efficiency, from the decarbonization of the supply of electricity and fuels (by means of fuel shifting, carbon capture and storage, nuclear energy, and renewable energy), and from biological storage in forests and agricultural soils. Below, we discuss 15 different examples of options that are already deployed at an industrial scale and that could be scaled up further to produce at least one wedge (summarized in Table 1). Although several options could be scaled up to two or more wedges, we doubt that any could fill the stabilization triangle, or even half of it, alone.

Because the same BAU carbon emissions cannot be displaced twice, achieving one wedge often interacts with achieving another. The more the electricity system becomes decarbonized, for example, the less the available savings from greater efficiency of electricity use, and vice versa. Interactions among wedges are discussed in the SOM text. Also, our focus is not on costs. In general, the achievement of a wedge will require some price trajectory for carbon, the details of which depend on many assumptions, including future fuels prices, public acceptance, and cost reductions by means of learning. Instead, our analysis is intended to complement the comprehensive but complex "integrated assessments" (*1*) of carbon mitigation by letting the full-scale examples that are already in the marketplace make a simple case for technological readiness.

Category I: Efficiency and Conservation

Improvements in efficiency and conservation probably offer the greatest potential to provide wedges. For example, in 2002, the United States announced the goal of decreasing its carbon intensity (carbon emissions per unit GDP) by 18% over the next decade, a decrease of 1.96% per year. An entire wedge would be created if the United States were to reset its carbon intensity goal to a decrease of 2.11% per year and extend it to 50 years, and if every country were to follow suit by adding the same 0.15% per year increment to its own carbon intensity goal. However, efficiency and conservation options are less tangible than those from the other categories. Improvements in energy efficiency will come from literally hundreds of innovations that range from new catalysts and chemical processes, to more efficient lighting and insulation for buildings, to the growth of the service economy and telecommuting. Here, we provide four of many possible comparisons of greater and less efficiency in 2054. (See references and details in Section 2 of the SOM text.)

Option 1: Improved fuel economy. Suppose that in 2054, 2 billion cars (roughly four times as many as today) average 10,000 miles per year (as they do today). One wedge would be achieved if, instead of averaging 30 miles per gallon (mpg) on conventional fuel, cars in 2054 averaged 60 mpg, with fuel type and distance traveled unchanged.

Option 2: Reduced reliance on cars. A wedge would also be achieved if the average fuel economy of the 2 billion 2054 cars were 30 mpg, but the annual distance traveled were 5000 miles instead of 10,000 miles.

Option 3: More efficient buildings. According to a 1996 study by the IPCC, a wedge is the difference between pursuing and not pursuing "known and established approaches" to energy-efficient space heating and cooling, water heating, lighting, and refrigeration in residential and commercial buildings. These approaches reduce mid-century emissions from buildings by about one-fourth. About half of potential savings are in the buildings in developing countries (*1*).

Option 4: Improved power plant efficiency. In 2000, coal power plants, operating on average at 32% efficiency, produced about one-fourth of all carbon emissions: 1.7 GtC/year out of 6.2 GtC/year. A wedge would be created if twice today's quantity of coal-based electricity in 2054 were produced at 60% instead of 40% efficiency.

Category II: Decarbonization of Electricity and Fuels

(See references and details in Section 3 of the SOM text.)

Option 5: Substituting natural gas for coal. Carbon emissions per unit of electricity are about half as large from natural gas power plants as from coal plants. Assume that the capacity factor of the average baseload coal plant in 2054 has increased to 90% and that its efficiency has improved to 50%. Because 700 GW of such plants emit carbon at a rate of 1 GtC/year, a wedge would be achieved by displacing 1400 GW of baseload coal with baseload gas by 2054. The power shifted to gas for this wedge is four times as large as the total current gas-based power.

Option 6: Storage of carbon captured in power plants. Carbon capture and storage (CCS) technology prevents about 90% of the fossil carbon from reaching the atmosphere, so a wedge would be provided by the installation of CCS at 800 GW of baseload coal plants by 2054 or 1600 GW of baseload natural gas plants. The most likely approach

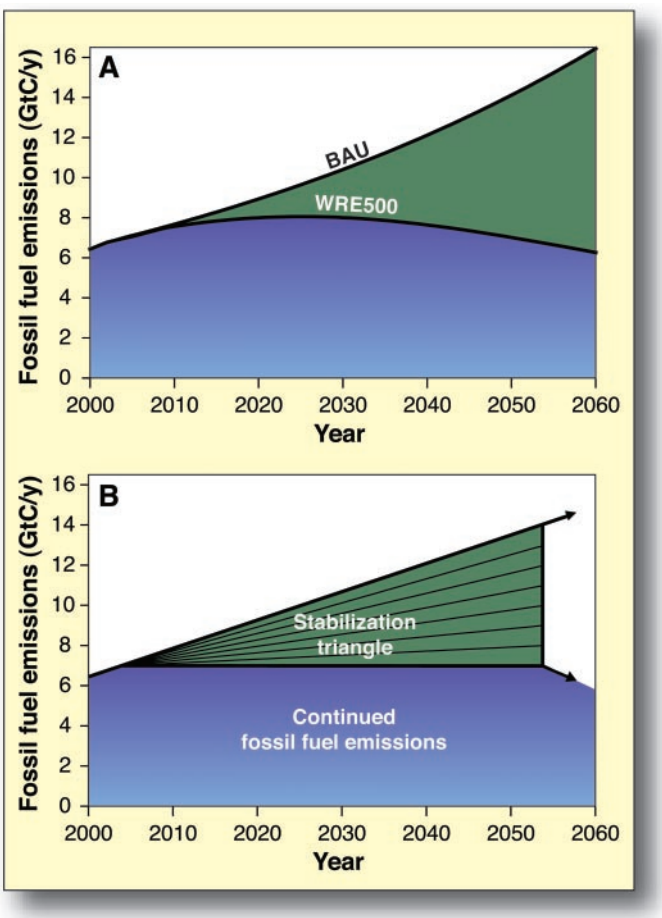


Fig. 1. (**A**) The top curve is a representative BAU emissions path for global carbon emissions as CO_2 from fossil fuel combustion and cement manufacture: 1.5% per year growth starting from 7.0 GtC/year in 2004. The bottom curve is a CO_2 emissions path consistent with atmospheric CO_2 stabilization at 500 ppm by 2125 akin to the Wigley, Richels, and Edmonds (WRE) family of stabilization curves described in (*11*), modified as described in Section 1 of the SOM text. The bottom curve assumes an ocean uptake calculated with the High-Latitude Exchange Interior Diffusion Advection (HILDA) ocean model (*12*) and a constant net land uptake of 0.5 GtC/year (Section 1 of the SOM text). The area between the two curves represents the avoided carbon emissions required for stabilization. (**B**) Idealization of (A): A stabilization triangle of avoided emissions (green) and allowed emissions (blue). The allowed emissions are fixed at 7 GtC/year beginning in 2004. The stabilization triangle is divided into seven wedges, each of which reaches 1 GtC/year in 2054. With linear growth, the total avoided emissions per wedge is 25 GtC, and the total area of the stabilization triangle is 175 GtC. The arrow at the bottom right of the stabilization triangle points downward to emphasize that fossil fuel emissions must decline substantially below 7 GtC/year after 2054 to achieve stabilization at 500 ppm.

has two steps: (i) precombustion capture of CO_2, in which hydrogen and CO_2 are produced and the hydrogen is then burned to produce electricity, followed by (ii) geologic storage, in which the waste CO_2 is injected into subsurface geologic reservoirs. Hydrogen production from fossil fuels is already a very large business. Globally, hydrogen plants consume about 2% of primary energy and emit 0.1 GtC/year of CO_2. The capture part of a wedge of CCS electricity would thus require only a tenfold expansion of plants resembling today's large hydrogen plants over the next 50 years.

The scale of the storage part of this wedge can be expressed as a multiple of the scale of current enhanced oil recovery, or current seasonal storage of natural gas, or the first geological storage demonstration project. Today, about 0.01 GtC/year of carbon as CO_2 is injected into geologic reservoirs to spur enhanced oil recovery, so a wedge of geologic storage requires that CO_2 injection be scaled up by a factor of 100 over the next 50 years. To smooth out seasonal demand in the United States, the natural gas industry annually draws roughly 4000 billion standard cubic feet (Bscf) into and out of geologic storage, and a carbon flow of 1 GtC/year (whether as methane or CO_2) is a flow of 69,000 Bscf/year (190 Bscf per day), so a wedge would be a flow to storage 15 and 20 times as large as the current flow. Norway's Sleipner project in the North Sea strips CO_2 from natural gas offshore and reinjects 0.3 million tons of carbon a year (MtC/year) into a non–fossil-fuel–bearing formation, so a wedge would be 3500 Sleipner-sized projects (or fewer, larger projects) over the next 50 years.

A worldwide effort is under way to assess the capacity available for multicentury storage and to assess risks of leaks large enough to endanger human or environmental health.

Option 7: Storage of carbon captured in hydrogen plants. The hydrogen resulting from precombustion capture of CO_2 can be sent offsite to displace the consumption of conventional fuels rather than being consumed onsite to produce electricity. The capture part of a wedge

Table 1. Potential wedges: Strategies available to reduce the carbon emission rate in 2054 by 1 GtC/year or to reduce carbon emissions from 2004 to 2054 by 25 GtC.

Option	Effort by 2054 for one wedge, relative to 14 GtC/year BAU	Comments, issues
	Energy efficiency and conservation	
Economy-wide carbon-intensity reduction (emissions/$GDP)	Increase reduction by additional 0.15% per year (e.g., increase U.S. goal of 1.96% reduction per year to 2.11% per year)	Can be tuned by carbon policy
1. Efficient vehicles	Increase fuel economy for 2 billion cars from 30 to 60 mpg	Car size, power
2. Reduced use of vehicles	Decrease car travel for 2 billion 30-mpg cars from 10,000 to 5000 miles per year	Urban design, mass transit, telecommuting
3. Efficient buildings	Cut carbon emissions by one-fourth in buildings and appliances projected for 2054	Weak incentives
4. Efficient baseload coal plants	Produce twice today's coal power output at 60% instead of 40% efficiency (compared with 32% today)	Advanced high-temperature materials
	Fuel shift	
5. Gas baseload power for coal baseload power	Replace 1400 GW 50%-efficient coal plants with gas plants (four times the current production of gas-based power)	Competing demands for natural gas
	CO_2 Capture and Storage (CCS)	
6. Capture CO_2 at baseload power plant	Introduce CCS at 800 GW coal or 1600 GW natural gas (compared with 1060 GW coal in 1999)	Technology already in use for H_2 production
7. Capture CO_2 at H_2 plant	Introduce CCS at plants producing 250 MtH_2/year from coal or 500 MtH_2/year from natural gas (compared with 40 MtH_2/year today from all sources)	H_2 safety, infrastructure
8. Capture CO_2 at coal-to-synfuels plant	Introduce CCS at synfuels plants producing 30 million barrels a day from coal (200 times Sasol), if half of feedstock carbon is available for capture	Increased CO_2 emissions, if synfuels are produced without CCS
Geological storage	Create 3500 Sleipners	Durable storage, successful permitting
	Nuclear fission	
9. Nuclear power for coal power	Add 700 GW (twice the current capacity)	Nuclear proliferation, terrorism, waste
	Renewable electricity and fuels	
10. Wind power for coal power	Add 2 million 1-MW-peak windmills (50 times the current capacity) "occupying" 30×10^6 ha, on land or offshore	Multiple uses of land because windmills are widely spaced
11. PV power for coal power	Add 2000 GW-peak PV (700 times the current capacity) on 2×10^6 ha	PV production cost
12. Wind H_2 in fuel-cell car for gasoline in hybrid car	Add 4 million 1-MW-peak windmills (100 times the current capacity)	H_2 safety, infrastructure
13. Biomass fuel for fossil fuel	Add 100 times the current Brazil or U.S. ethanol production, with the use of 250×10^6 ha (one-sixth of world cropland)	Biodiversity, competing land use
	Forests and agricultural soils	
14. Reduced deforestation, plus reforestation, afforestation, and new plantations.	Decrease tropical deforestation to zero instead of 0.5 GtC/year, and establish 300 Mha of new tree plantations (twice the current rate)	Land demands of agriculture, benefits to biodiversity from reduced deforestation
15. Conservation tillage	Apply to all cropland (10 times the current usage)	Reversibility, verification

would require the installation of CCS, by 2054, at coal plants producing 250 MtH_2/year, or at natural gas plants producing 500 MtH_2/year. The former is six times the current rate of hydrogen production. The storage part of this option is the same as in Option 6.

Option 8: Storage of carbon captured in synfuels plants. Looming over carbon management in 2054 is the possibility of large-scale production of synthetic fuel (synfuel) from coal. Carbon emissions, however, need not exceed those associated with fuel refined from crude oil if synfuels production is accompanied by CCS. Assuming that half of the carbon entering a 2054 synfuels plant leaves as fuel but the other half can be captured as CO_2, the capture part of a wedge in 2054 would be the difference between capturing and venting the CO_2 from coal synfuels plants producing 30 million barrels of synfuels per day. (The flow of carbon in 24 million barrels per day of crude oil is 1 GtC/year; we assume the same value for the flow in synfuels and allow for imperfect capture.) Currently, the Sasol plants in South Africa, the world's largest synfuels facility, produce 165,000 barrels per day from coal. Thus, a wedge requires 200 Sasol-scale coal-to-synfuels facilities with CCS in 2054. The storage part of this option is again the same as in Option 6.

Option 9: Nuclear fission. On the basis of the Option 5 estimates, a wedge of nuclear electricity would displace 700 GW of efficient baseload coal capacity in 2054. This would require 700 GW of nuclear power with the same 90% capacity factor assumed for the coal plants, or about twice the nuclear capacity currently deployed. The global pace of nuclear power plant construction from 1975 to 1990 would yield a wedge, if it continued for 50 years (*10*). Substantial expansion in nuclear power requires restoration of public confidence in safety and waste disposal, and international security agreements governing uranium enrichment and plutonium recycling.

Option 10: Wind electricity. We account for the intermittent output of windmills by equating 3 GW of nominal peak capacity (3 GW_p) with 1 GW of baseload capacity. Thus, a wedge of wind electricity would require the deployment of 2000 GW_p that displaces coal electricity in 2054 (or 2 million 1-MW_p wind turbines). Installed wind capacity has been growing at about 30% per year for more than 10 years and is currently about 40 GW_p. A wedge of wind electricity would thus require 50 times today's deployment. The wind turbines would "occupy" about 30 million hectares (about 3% of the area of the United States), some on land and some offshore. Because windmills are widely spaced, land with windmills can have multiple uses.

Option 11: Photovoltaic electricity. Similar to a wedge of wind electricity, a wedge from photovoltaic (PV) electricity would require 2000 GW_p of installed capacity that displaces coal electricity in 2054. Although only 3 GW_p of PV are currently installed, PV electricity has been growing at a rate of 30% per year. A wedge of PV electricity would require 700 times today's deployment, and about 2 million hectares of land in 2054, or 2 to 3 m^2 per person.

Option 12: Renewable hydrogen. Renewable electricity can produce carbon-free hydrogen for vehicle fuel by the electrolysis of water. The hydrogen produced by 4 million 1-MW_p windmills in 2054, if used in high-efficiency fuel-cell cars, would achieve a wedge of displaced gasoline or diesel fuel. Compared with Option 10, this is twice as many 1-MW_p windmills as would be required to produce the electricity that achieves a wedge by displacing high-efficiency baseload coal. This interesting factor-of-two carbon-saving advantage of wind-electricity over wind-hydrogen is still larger if the coal plant is less efficient or the fuel-cell vehicle is less spectacular.

Option 13: Biofuels. Fossil-carbon fuels can also be replaced by biofuels such as ethanol. A wedge of biofuel would be achieved by the production of about 34 million barrels per day of ethanol in 2054 that could displace gasoline, provided the ethanol itself were fossil-carbon free. This ethanol production rate would be about 50 times larger than today's global production rate, almost all of which can be attributed to Brazilian sugarcane and United States corn. An ethanol wedge would require 250 million hectares committed to high-yield (15 dry tons/hectare) plantations by 2054, an area equal to about one-sixth of the world's cropland. An even larger area would be required to the extent that the biofuels require fossil-carbon inputs. Because land suitable for annually harvested biofuels crops is also often suitable for conventional agriculture, biofuels production could compromise agricultural productivity.

Category III: Natural Sinks

Although the literature on biological sequestration includes a diverse array of options and some very large estimates of the global potential, here we restrict our attention to the pair of options that are already implemented at large scale and that could be scaled up to a wedge or more without a lot of new research. (See Section 4 of the SOM text for references and details.)

Option 14: Forest management. Conservative assumptions lead to the conclusion that at least one wedge would be available from reduced tropical deforestation and the management of temperate and tropical forests. At least one half-wedge would be created if the current rate of clear-cutting of primary tropical forest were reduced to zero over 50 years instead of being halved. A second half-wedge would be created by reforesting or afforesting approximately 250 million hectares in the tropics or 400 million hectares in the temperate zone (current areas of tropical and temperate forests are 1500 and 700 million hectares, respectively). A third half-wedge would be created by establishing approximately 300 million hectares of plantations on nonforested land.

Option 15: Agricultural soils management. When forest or natural grassland is converted to cropland, up to one-half of the soil carbon is lost, primarily because annual tilling increases the rate of decomposition by aerating undecomposed organic matter. About 55 GtC, or two wedges' worth, has been lost historically in this way. Practices such as conservation tillage (e.g., seeds are drilled into the soil without plowing), the use of cover crops, and erosion control can reverse the losses. By 1995, conservation tillage practices had been adopted on 110 million hectares of the world's 1600 million hectares of cropland. If conservation tillage could be extended to all cropland, accompanied by a verification program that enforces the adoption of soil conservation practices that actually work as advertised, a good case could be made for the IPCC's estimate that an additional half to one wedge could be stored in this way.

Conclusions

In confronting the problem of greenhouse warming, the choice today is between action and delay. Here, we presented a part of the case for action by identifying a set of options that have the capacity to provide the seven stabilization wedges and solve the climate problem for the next half-century. None of the options is a pipe dream or an unproven idea. Today, one can buy electricity from a wind turbine, PV array, gas turbine, or nuclear power plant. One can buy hydrogen produced with the chemistry of carbon capture, biofuel to power one's car, and hundreds of devices that improve energy efficiency. One can visit tropical forests where clear-cutting has ceased, farms practicing conservation tillage, and facilities that inject carbon into geologic reservoirs. Every one of these options is already implemented at an industrial scale and could be scaled up further over 50 years to provide at least one wedge.

References and Notes

1. IPCC, *Climate Change 2001: Mitigation*, B. Metz *et al.*, Eds. (IPCC Secretariat, Geneva, Switzerland, 2001); available at www.grida.no/climate/ipcc_tar/wg3/index.htm.
2. M. I. Hoffert *et al.*, *Science* **298**, 981 (2002).
3. R. T. Watson *et al.*, *Climate Change 2001: Synthesis Report. Contribution to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, Cambridge, UK, 2001).
4. B. C. O'Neill, M. Oppenheimer, *Science* **296**, 1971 (2002).
5. Royal Commission on Environmental Pollution, *En-*

ergy: The Changing Climate (2000); available at www.rcep.org.uk/energy.htm.
6. Environmental Defense, *Adequacy of Commitments—Avoiding "Dangerous" Climate Change: A Narrow Time Window for Reductions and a Steep Price for Delay* (2002); available at www.environmentaldefense.org/documents/2422_COP_time.pdf.
7. "Climate OptiOns for the Long Term (COOL) synthesis report," *NRP Rep. 954281* (2002); available at www.wau.nl/cool/reports/COOLVolumeAdef.pdf.
8. IPCC, *Special Report on Emissions Scenarios* (2001); available at www.grida.no/climate/ipcc/emission/index.htm.
9. R. Socolow, S. Pacala, J. Greenblatt, Proceedings of the *Seventh International Conference on Greenhouse Gas Control Technology*, Vancouver, Canada, 5 to 9 September, 2004, in press.
10. BP, *Statistical Review of World Energy* (2003); available at www.bp.com/subsection.do?categoryId=95&contentId=2006480.
11. T. M. L. Wigley, in *The Carbon Cycle*, T. M. L. Wigley, D. S. Schimel, Eds. (Cambridge Univ. Press, Cambridge, 2000), pp. 258–276.
12. G. Shaffer, J. L. Sarmiento, *J. Geophys. Res.* **100**, 2659 (1995).
13. The authors thank J. Greenblatt, R. Hotinski, and R. Williams at Princeton; K. Keller at Penn State; and C. Mottershead at BP. This paper is a product of the Carbon Mitigation Initiative (CMI) of the Princeton Environmental Institute at Princeton University. CMI (www.princeton.edu/~cmi) is sponsored by BP and Ford.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/968/DC1
SOM Text
Figs. S1 and S2
Tables S1 to S5
References

VIEWPOINT

Sustainable Hydrogen Production

John A. Turner

Identifying and building a sustainable energy system are perhaps two of the most critical issues that today's society must address. Replacing our current energy carrier mix with a sustainable fuel is one of the key pieces in that system. Hydrogen as an energy carrier, primarily derived from water, can address issues of sustainability, environmental emissions, and energy security. Issues relating to hydrogen production pathways are addressed here. Future energy systems require money and energy to build. Given that the United States has a finite supply of both, hard decisions must be made about the path forward, and this path must be followed with a sustained and focused effort.

In his 2003 State of the Union Address, U.S. President Bush proposed "$1.2 billion in research funding so that America can lead the world in developing clean, hydrogen-powered automobiles." Since that time, articles both pro and con have buffeted the whole concept. The hydrogen economy (*1*) is not a new idea. In 1874, Jules Verne, recognizing the finite supply of coal and the possibilities of hydrogen derived from water electrolysis, made the comment that "water will be the coal of the future" (*2*). Rudolf Erren in the 1930s suggested using hydrogen produced from water electrolysis as a transportation fuel (*3*). His goal was to reduce automotive emissions and oil imports into England. Similarly, Francis Bacon suggested using hydrogen as an energy storage system (*4*). The vision of using energy from electricity and electrolysis to generate hydrogen from water for transportation and energy storage to reduce environmental emissions and provide energy security is compelling, but as yet remains unrealized.

If one assumes a full build-out of a hydrogen economy, the amount of hydrogen needed just for U.S. transportation needs would be about 150 million tons per year (*5*). One must question the efficacy of producing, storing, and distributing that much hydrogen. Because energy is required to extract hydrogen from either water or biomass so that it can be used as an energy carrier, if the United States chooses a hydrogen-based future it needs to think carefully about how much energy we need and where it is going to come from. In addition, sustainability must be a hallmark of any proposed future infrastructure. What energy-producing technologies can be envisioned that will last for millennia, and just how many people can they support (*6*–*8*)?

National Renewable Energy Laboratory, Golden, CO 80401–3393, USA. E-mail: jturner@nrel.gov

Technologies for Hydrogen Production

Hydrogen can be generated from water, biomass, natural gas, or (after gasification) coal. Today, hydrogen is mainly produced from natural gas via steam methane reforming, and although this process can sustain an initial foray into the hydrogen economy, it represents only a modest reduction in vehicle emissions as compared to emissions from current hybrid vehicles, and ultimately only exchanges oil imports for natural gas imports. It is clearly not sustainable.

Coal gasification could produce considerable amounts of hydrogen and electricity merely because of the large size of available coal deposits (*9*). Additionally, because of its relatively low cost, it is often cited as the best resource for economically producing large quantities of hydrogen. However, the energy required for the necessary sequestration of CO_2 would increase the rate at which coal reserves are depleted; converting the vehicle fleet to electric vehicles and generating that electricity from "clean coal" or making hydrogen as a possible energy carrier would accelerate that depletion. Couple that to a modest economic growth rate of ~1%, and U.S. 250-year coal reserves drop to 75 years or so (*6*), which is not at all sustainable. That leaves solar-derived, wind, nuclear, and geothermal energy as major resources for sustainable hydrogen production. The hydrogen production pathways from these resources include electrolysis of water, thermal chemical cycles using heat, and biomass processing (using a variety of technologies ranging from reforming to fermentation).

Biomass processing techniques can benefit greatly from the wealth of research that has been carried out over the years on refining and converting liquid and gaseous fossil fuels. Some of these processes require considerable amounts of hydrogen, and many of these fossil-derived processes can be adapted for use with a large variety of biomass-derived feedstocks. Biomass can easily be converted into a number of liquid fuels, including methanol, ethanol, biodiesel, and pyrolysis oil, which could be transported and used to generate hydrogen on site. For the high-biomass-yield processes, such as corn to ethanol, hydrogen is required in the form of ammonia for fertilizer. Although biomass is clearly (and necessarily) sustainable, it cannot supply hydrogen in the amounts required. It remains to be seen, in a world that is both food-limited and carbon-constrained, whether the best use of biomass is for food, as a chemical feedstock, or as an energy source.

Because the direct thermal splitting of water requires temperatures of >2000°C and produces a rapidly recombining mixture of hydrogen and oxygen (*10*), a number of thermal chemical cycles have been identified that can use lower temperatures and produce hydrogen and oxygen in separate steps. The one that has received the greatest attention involves sulfuric acid (H_2SO_4) at 850°C and hydrogen iodide (HI) at 450°C (*11*). The next generation of fission reactors includes designs that can provide the necessary heat; however, a number of critical material properties must be satisfied to meet the required stability under the operating conditions of HI

and H_2SO_4. For safety reasons, a fairly long heat transfer line (~1 km) is necessary, so that the hydrogen-producing chemical plant is located away from the reactor. If the issues of nuclear proliferation and reprocessing can be dealt with, then reactors based on these designs could potentially supply many hundreds of years of energy, but even that is not ultimately sustainable. Solar thermal systems could also be used to drive such thermal chemical cycles, although more interesting cycles involve the use of metal/metal oxide systems, in which solar heat is used to convert an oxide to the metal (releasing oxygen), and then the metal is reacted with water to produce hydrogen and reform the oxide (*12*).

Any technology that produces electricity can drive an electrolyzer to produce hydrogen. Because of the enormous potential of solar and wind (*13*), it seems possible that electrolysis can supply future societies with whatever hydrogen would be necessary. Figure 1 shows the cost of hydrogen from electrolysis, based on the cost of the electricity and the efficiency of the electrolyzer (note that these are system efficiencies and include all losses) (*14*, *15*). For example, some systems provide high-pressure (70-MPa) hydrogen via electrochemical pressurization. Average U.S. electricity prices range from 4.8¢ for large-scale industrial users to 8.45¢ for commercial users (*16*). Based on thermodynamic considerations alone, improvements in the efficiency of electrolysis are not going to lead to major reductions in the cost of produced hydrogen. Additionally, as the cost of electricity goes down [unsubsidized wind is already below 4¢ per kilowatt-hour (kWh)], efficiency has a lower impact on the cost of the hydrogen. Rather, improvements and innovations in the capital cost of the plant and the lifetime of the cell and its maintenance requirements are where the major cost savings will likely be obtained.

System efficiencies of commercial electrolyzers range from 60 to 73%, so one argument often used to discount electrolysis is its perceived low efficiency. However, although efficiency is certainly important, it is neither a good proxy for deciding on new technology, nor should it be the determining factor. If combined-cycle natural gas plants had the same efficiency as coal plants, they wouldn't be economical at all; and even with their higher efficiency, they produce electricity at a higher cost than coal.

The energy required to split water can be obtained from a combination of heat and electricity. At 25°C, there is enough heat in the environment that the electricity requirement drops to 1.23 V. Increasing the electrolysis temperature can lower the electrolysis voltage, but the total amount of energy required to split water remains relatively constant (actually, the isothermal potential increases slightly). Thus, higher-temperature electrolysis only makes sense if the heat is free and it only requires a small amount of energy to move it where you need it, or there is an advantage in a new material set (lower cost, longer lifetime, etc.) or a significant decrease in the electrolysis energy losses. Possible areas for heat plus electrolysis options include nuclear, geothermal, and a number of solar-based configurations.

The amount of water needed to produce hydrogen for transportation is not great. Conversion of the current U.S. light-duty fleet (some 230 million vehicles) to fuel cell vehicles would require about 100 billion gallons of water/year to supply the needed hydrogen (*17*). Domestic personal water use in the United States is about 4800 billion gallons/year. The U.S. uses about 300 billion gallons of water/year for the production of gasoline (*18*), and about 70 trillion gallons of water/year for thermoelectric power generation (*19*). Solar and wind power do not require water for their electricity generation. So not only do these resources provide sustainable carbon-free energy, they reduce the water requirements for power generation.

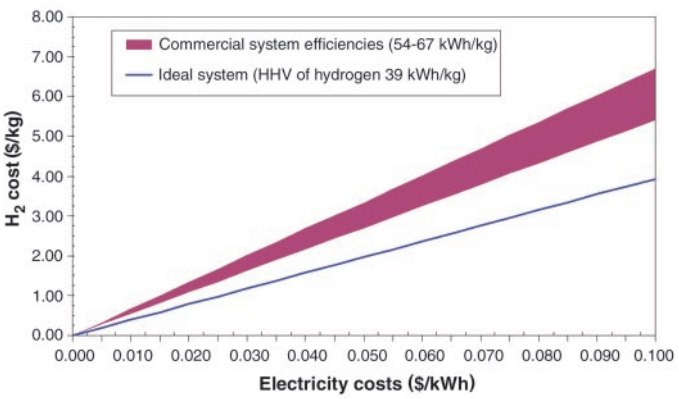


Fig. 1. The cost of hydrogen based on the electricity prices alone; no capital, operating, or maintenance costs are included in the calculation. HHV, higher heating value.

Impurities in the water can significantly reduce the lifetime of the electrolysis cell. Water is usually purified on site, but water cleanup could add to the cost of the hydrogen. In a stationary system where hydrogen is used for energy storage, the water from the fuel cell could be cycled back to the electrolyzer with minimal purification.

Sustainable hydrogen production technologies that may affect hydrogen production in the future include photobiological (*20*) and photoelectrochemical approaches (*21*–*23*). These systems produce hydrogen directly from sunlight and water, and offer the possibility of increasing the efficiency of the solar-to-hydrogen pathway (*24*) and lowering the capital cost of the system, but they still require land area to collect sunlight. These systems might allow the use of seawater directly as the feedstock instead of high-purity water.

General Comments

An important consideration is the energy payback during a time of rapid growth of a new energy or energy carrier technology. There will likely be an extended period of time when the new technologies consume more energy than they produce. The time frame for conversion to an alternative energy system is typically/historically 75 to 100 years. With this in mind, we need to think carefully about how many intermediate technology steps we introduce and how long (and at what cost) we must operate them in order to make the energy payback positive. The energy required to sustain a growth rate must also be taken into account.

Most hydrogen-producing systems being proposed are smaller than the current centralized power plants. Instead of building a small number of large generating plants, a large number of smaller plants such as wind farms and solar arrays are proposed that, when added together, can produce large amounts of energy. To be considered then is the benefit of a technology that is amenable to mass manufacturing. Much higher volumes can translate into cost savings. Electrolyzers, fuel cells, and battery technologies all fall into this area.

Although a great deal of money, thought, and energy are currently going into sequestration technologies, the question still remains: Is this the best way to spend our limited supply of energy and financial capital? As I said earlier, the best use of carbon-free sustainable electricity would be to replace coal-burning power plants (*13*). Just because we have large coal reserves does not mean that we must use them. The question is whether we have the will to leave that energy in the ground and move on to something more advanced. Sustainable energy systems can easily provide (albeit at some cost) sufficient amounts of both electricity and hydrogen. Although current gasoline-powered hybrid vehicles can reduce fossil fuel use, they cannot eliminate it. For transportation, the research, development, and demonstration of the hydrogen economy are well served by using the existing natural gas–based infra-

structure. Integrating sustainable energy systems into the infrastructure would allow rapid adoption of electrolysis-based hydrogen production, whenever these future transportation systems become viable. Since the 1930s, the recognized vision of the hydrogen economy has been to allow the storage of electrical energy, reduce environmental emissions, and provide a transportation fuel. This goal is clearly achievable, but only with a sustained, focused effort.

References and Notes

1. For the purpose of this discussion, I use the following definition of the hydrogen economy: the production, storage, distribution, and use of hydrogen as an energy carrier.
2. Jules Verne, *The Mysterious Island* (available at http://www.literature-web.net/verne/mysteriousisland, 1874).
3. P. Hoffmann, *The Forever Fuel: The Story of Hydrogen* (Westview Press, Boulder, CO, 1981).
4. D. Gregory, *Sci. Am.* **228**, (no. 1), 13 (January 1973).
5. *Basic Research Needs for the Hydrogen Economy*, available at http://www.sc.doe.gov/bes/reports/files/NHE_rpt.pdf (current U.S. production is about 9 million tons of hydrogen per year).
6. P. Weisz, *Phys. Today* **57** (no. 7), 47 (2004).
7. A. Bartlett, *Phys. Today* **57** (no. 7), 53 (2004).
8. G. Richard, *ILEA Leaf*, Winter 2002 (available at www.ilea.org/leaf/richard2002.html).
9. Energy Information Administration, unpublished file data of the Coal Reserves Data Base (February 2004), available at http://www.eia.doe.gov/pub/international/iea2002/table82.xls.
10. A. Steinfeld, *Solar Energy*, in press (available online 3 February 2004).
11. *Nuclear Production of Hydrogen, Second Information Exchange Meeting–Argonne, Illinois, USA 2-3 October 2003* (Organisation for Economic Cooperation and Development, Paris) (available at http://oecdpublications.gfi-nb.com/cgi-bin/OECDBookShop.storefront/EN/product/662004021P1).
12. A. Steinfeld, *Int. J. Hydrogen Energy* **27**, 611 (2002).
13. J. A. Turner, *Science* **285**, 5428 (1999).
14. J. Ivy, *Summary of Electrolytic Hydrogen Production: Milestone Completion Report*, available at www.osti.gov/servlets/purl/15007167-aF4rPu/native/.
15. In any discussion concerning the efficiency of electrolyzers, it is appropriate to use the higher heating value to calculate the efficiency. This corresponds to the isothermal potential (1.47 V = 39 kWh/kg) and represents the assumption that all the energy needed to split water comes from the electricity.
16. These figures are from the Energy Information Administration, available at www.eia.doe.gov/cneaf/electricity/epm/tablees1a.html.
17. For an estimate of the amount of water needed for hydrogen-powered fuel cell vehicles, we will assume a vehicle fuel economy of 60 miles per kg of H_2, that vehicle miles traveled = 2.6×10^{12} miles/year (found at www.bts.gov/publications/national_transportation_statistics/2002/html/table_automobile_profile.html), and that 1 gallon of water contains 0.42 kg of H_2. Total water required for the U.S. fleet = (2.6×10^{12} miles/year)(1 kg of H_2/60 miles)(1 gal H_2O/0.42 kg of H_2) = 1.0×10^{11} gallons of H_2O/year. This represents the water used directly for fuel. If one considers all water uses along the chain; for example, from construction of wind farms to the electrolysis systems (life cycle assessment), then the total water use would be in the range of 3.3×10^{11}gallons H_2O/year.
18. This is a life cycle analysis (M. Mann and M. Whitaker, unpublished data). The United States used about 126 billion gallons of gasoline in 2001 [see link in (17)].
19. See http://water.usgs.gov/pubs/circ/2004/circ1268/.
20. A. Melis, *Int. J. Hydrogen Energy* **27**, 1217 (2002).
21. O. Khaselev, J. A. Turner, *Science* **280**, 425 (1998).
22. M. Graetzel, *Nature* **414**, 338 (2001).
23. N. Lewis, *Nature* **414**, 589 (2001).
24. O. Khaselev, A. Bansal, J. A. Turner, *Int. J. Hydrogen Energy* **26**, 127 (2001).
25. Contributions by D. Sandor for careful manuscript edits and by J. Ivy for Fig. 1 are gratefully acknowledged.

VIEWPOINT

Hybrid Cars Now, Fuel Cell Cars Later

Nurettin Demirdöven[1] and John Deutch[2]*

We compare the energy efficiency of hybrid and fuel cell vehicles as well as conventional internal combustion engines. Our analysis indicates that fuel cell vehicles using hydrogen from fossil fuels offer no significant energy efficiency advantage over hybrid vehicles operating in an urban drive cycle. We conclude that priority should be placed on hybrid vehicles by industry and government.

Our interest in moving toward a hydrogen economy has its basis not in love of the molecule but in the prospect of meeting energy needs at acceptable cost, with greater efficiency and less environmental damage compared to the use of conventional fuels. One goal is the replacement of today's automobile with a dramatically more energy-efficient vehicle. This will reduce carbon dioxide emissions that cause adverse climate change as well as dependence on imported oil. In 2001, the United States consumed 8.55 million barrels of motor gasoline per day (*1*), of which an estimated 63.4% is refined from imported crude oil (*2*). This consumption resulted in annual emissions of 308 million metric tons (MMT) of carbon equivalent in 2001, accounting for 16% of total U.S. carbon emissions of 1892 MMT (*3*).

Two advanced vehicle technologies that are being considered to replace the current fleet, at least partially, are hybrid vehicles and fuel cell (FC)–powered vehicles. Hybrid vehicles add a parallel direct electric drive train with motor and batteries to the conventional internal combustion engine (ICE) drive train. This hybrid drive train permits significant reduction in idling losses and regeneration of braking losses that leads to greater efficiency and improved fuel economy. Hybrid technology is available now, although it represents less than 1% of new car sales. FC vehicles also operate by direct current electric drive. They use the high efficiency of electrochemical fuel cells to produce power from hydrogen. For the foreseeable future, hydrogen will come from fossil fuels by reforming natural gas or gasoline. FC vehicle technology is not here today, and commercialization will require a large investment in research, development, and infrastructure (*4*).

Here, we evaluate the potential of these advanced passenger vehicles to improve energy efficiency. We show that a tremendous increase in energy efficiency can be realized today by shifting to hybrid ICE vehicles, quite likely more than can be realized by a shift from hybrid ICE to hybrid FC vehicles.

Energy Efficiency Model

To provide a basis for comparison of these two technologies, we use a simple model (*5*) for obtaining the energy efficiency of the various power plant–drive train–fuel combinations considered in more detailed studies (*6*–*11*). In general, the energy efficiency of ICEs with a hybrid drive train and from FC-powered vehicles vary depending on the vehicle configuration and the type of engine, drive train, and fuel (natural gas, gasoline, or diesel).

For each configuration, we determine well-to-wheel (WTW) energy efficiency for a vehicle of a given weight operating on a specified drive cycle. The overall WTW efficiency is divided into a well-to-tank (WTT) and tank-to-wheel (TTW) efficiency so that WTW = WTT × TTW.

We begin with the U.S. Department of Energy (DOE) specification of average passenger energy use in a federal urban drive cycle, the so-called FUDS cycle (*12*). For example, for today's ICE vehicle that uses a spark ignition engine fueled by gasoline, the TTW efficiency for propulsion and braking is 12.6% (Fig. 1A).

[1]Technology and Policy Program, Engineering Systems Division, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. J.D. is a director of Cummins (a manufacturer of diesel engines) and an advisor to United Technologies Corporation, UTC Power, a fuel cell manufacturer.

*To whom correspondence should be addressed. E-mail: jmd@mit.edu

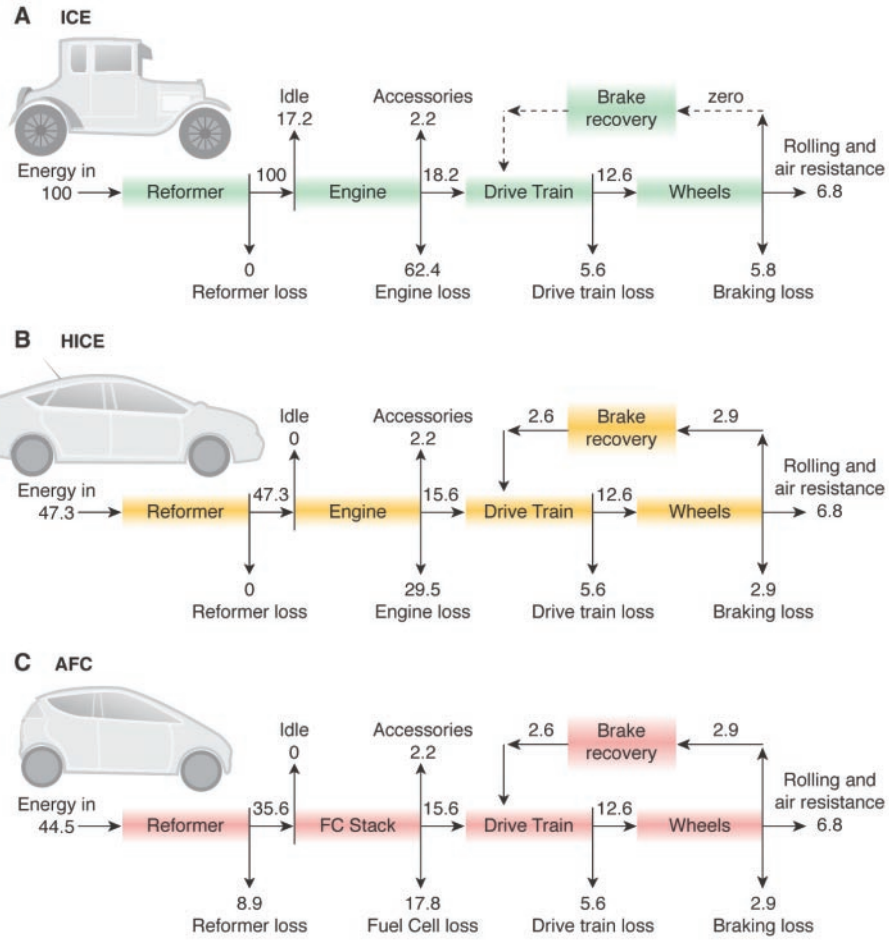


Fig. 1. Energy flow for various vehicle configurations. (**A**) ICE, the conventional internal combustion, spark ignition engine; (**B**) HICE, a hybrid vehicle that includes an electric motor and parallel drive train which eliminates idling loss and captures some energy of braking; (**C**) AFC a fuel cell vehicle with parallel drive train. The configuration assumes on-board gasoline reforming to fuel suitable for PEM fuel cell operation.

The TTW efficiency of other configurations is estimated by making changes in the baseline ICE parameters and calculating energy requirements beginning with energy output. A hypothetical hybrid ICE (HICE), based on current hybrid technology, that completely eliminates idling losses and captures a portion (50%) of braking losses for productive use (*13*) will have a TTW efficiency of 26.6% (Fig. 1B). Both the ICE and the HICE use gasoline fuel directly, so no fuel processing is needed.

A likely hydrogen-based car might be a proton-exchange membrane (PEM) FC-powered vehicle with a hybrid power train. This advanced fuel cell (AFC) vehicle has an on-board fuel processor that reforms gasoline to hydrogen fuel suitable for feed for the PEM fuel cell. We assume a reformer efficiency of 80% and 50% efficiency for the FC stack operating over the urban drive cycle. We include a power train with the same characteristics as the HICE vehicle. The TTW efficiency of this configuration is 28.3% (Fig. 1C).

It is apparent that any alternative vehicle configuration of fuel–power plant–drive train can be considered in a similar fashion. For example, if hydrogen were available without energy cost, the overall efficiency would improve to 39.0%, over three times that for the conventional ICE (*14*). A diesel ICE with a hybrid power train could achieve an efficiency of 31.9%, assuming that this higher compression direct-injection engine has an efficiency of 45.0% compared to 37.6% for the gasoline ICE.

Our results (Fig. 2) are in reasonably good agreement with those of more detailed studies but do not require elaborate simulation models. Figure 2 shows that, except for the Argonne National Laboratory/General Motors (ANL/GM) (*6*) study, the relative gain in efficiency in moving from an ordinary ICE to a HICE is more than twofold. The reason for this difference is not clear, because the TTW analysis in that study has its basis in a GM proprietary simulation model.

Validating Our Model

To test the validity of these comparisons and our simple model, we have used an advanced vehicle simulator called ADVISOR, developed by the National Renewable Research Laboratory (NREL) of DOE (*15*). ADVISOR provides estimates of energy efficiencies for different vehicle configurations. ADVISOR shows the broad range of vehicle performance that is possible with a reasonable choice of system parameters such as maximum engine power, maximum motor power, transmission type, and brake energy regeneration. The parameters we selected for the simulation of the ICE, HICE, and AFC are given in Table 1; for comparison, TTWs based on this simulation are 28.8% for the Toyota Prius and 26.2% for the Honda Insight. Except for the ANL/GM results, all studies point to large potential energy efficiency gains from hybrid vehicles in urban drive cycles compared to cars with conventional ICEs (*16*).

Our analysis shows that hybrids offer the potential for tremendous improvement in energy use and significant reduction in carbon emissions compared to current ICE technology. But hybrid vehicles will only be adopted in significant quantities if the cost to the consumer is comparable to the conventional ICE alternative. Hybrid technology is here today, but, of course, hybrid vehicles cost more than equivalent ICE vehicles because of the parallel drive train. Estimates of the cost differential vary, but a range of $1000 to $2000 is not unreasonable. Depending on the miles driven, the cost of ownership of a hybrid vehicle may be lower than a conventional ICE, because the discounted value of the fuel saving is greater than the incremental capital cost for the parallel drive train and

MODEL	ICE	HICE	AFC
Simple model	11.3	23.9	25.5
MIT-LFEE 2000	11.7	23.8	23.8-28.4
ANL/GM	15.2	18.6	24.6
NREL ADVISOR	11.3	24.5	23.9

Fig. 2. Comparison of WTW energy efficiencies of advanced vehicle systems using gasoline fuel. Color coding follows that in Fig. 1. 90% WTT efficiency in all cases; thus WTW = 0.90 TTW. Data for ICE and HICE is from (*7*), table 5.3. Data for AFC is from (*8*), which does not give energy efficiency directly. We derive a range for energy efficiency by comparing data in tables 8 and 9 for MJ/km for vehicle and fuel cycle for the 2020 ICE hybrid to that of the gasoline FC hybrid given in (*7*), table 5.3. Data from (*6*), table 2.1. Data from NREL's ADVISOR simulation; for details, see Table 1.

Table 1. Input and output vehicle parameters obtained from NREL's ADVISOR simulations. We assumed 1500 kg for the total vehicle weight, including two passengers and fuel on board. The actual weights of the Toyota Prius and Honda Insight with two passengers and fuel on board are 1368 kg and 1000 kg, respectively. Auxiliary power is 700 W except for the Honda Insight, for which it is 200 W. The simulations are over a FUDS cycle. Fuel use and TTW calculations follow the definition of efficiency given in (*5*), which is different than the "overall system efficiency" defined in the NREL's ADVISOR. Of course, the underlying performance is the same.

	ICE	HICE	AFC	Prius	Insight
			Vehicle		
Max power (kW)	102	83	70	74	60
Power:weight ratio (W/kg)	68	55	47	54	60
Frontal area (m^2)	2	2	2	1.75	1.9
Rolling resistance coefficient	0.009	0.009	0.009	0.009	0.0054
			Engine–motor–fuel cell stack		
Max engine power (kW)	102	43		43	50
Max engine efficiency (%)	38	38		39	40
Max motor power (kW)		40	40	31	10
Max motor efficiency (%)		92	92	91	96
Max fuel cell power (kW)			30		
Max fuel cell stack efficiency (%)			56		
			Acceleration		
Time for 0 to 60 mph (s)	18	10	13	15	12
			Fuel use		
Fuel energy use (kJ/km)	3282	1536	1553	1317 (1274)	1189 (982)
Fuel economy (mpg)	21	44	43	53	69
			Average efficiencies (%)		
Engine efficiency	21	30		28	25
Motor efficiency		79	84	81	90
Reformer efficiency			80		
Fuel cell stack efficiency			51		
Round-trip battery efficiency		100	84	81	82
Transmission efficiency	75	75	93	100	92
Regenerative braking efficiency		35	39	41	38
TTW efficiency	12.6	27.2	26.6	28.8	26.2

electric motor. Thus, hybrid vehicles can con tribute to lower emissions and less petroleum use at small or negative social cost (*17*). Today only Toyota and Honda offer hybrids in the United States; Daimler-Chrysler, Ford, and General Motors are planning to introduce hybrids in the period from 2004 to 2006. At present there is a federal tax credit of $1500 for purchase of a hybrid vehicle, but it is scheduled to phase out in 2006 (*18*).

Fuel cell technology is not here today. Both the Bush Administration's FreedomCAR program and the earlier Clinton Administration Partnership for a New Generation of Vehicles (PNGV) launched major DOE research and development initiatives for FC-powered vehicles. The current FreedomCAR program "focuses government support on fundamental, high-risk research that applies to multiple passenger-vehicle models and emphasizes the development of fuel cells and hydrogen infrastructure technologies" (*19*). A successful automotive FC program must develop high-durability FC stacks with lifetimes of 5 to 10 thousand hours, well beyond today's experience. It is impossible to estimate today whether the manufacturing cost range that FC stacks must achieve for economical passenger cars can be reached even at the large-scale production runs that might be envisioned.

The government FC research and development initiative is welcome, but it is not clear whether the effort to develop economic FC power plants for passenger cars will be successful. In parallel, we should place priority on deploying hybrid cars, beginning with today's automotive platforms and fuels. If the justification for federal support for research and development on fuel cells is reduction in imported oil and carbon dioxide emissions, then there is stronger justification for federal support for hybrid vehicles that will achieve similar results more quickly. Consideration should be given to expanding government support for research and development on generic advanced hybrid technology and extending hybrid vehicle tax credits.

References and Notes

1. Calculated from weekly data of supplied gasoline products published by DOE, Energy Information Agency; see www.eia.doe.gov/oil_gas/petroleum/info_glance/gasoline.html.
2. "National transportation statistics 2002," U.S. Department of Transportation, Bureau of Transportation Statistics (BTS02-08, Washington, DC, 2002), table 4-1.
3. "Inventory of U.S. greenhouse gas emissions and sinks: 1990-2001," final version, U.S. Environmental Protection Agency (EPA 430-R-03-004, Washington, DC, 2003), table A-1.
4. Only gasoline and natural gas are widely available as a transportation fuel today; a hydrogen or methanol fueled transportation system would take decades to deploy, at significant cost.
5. We define the average energy efficiency as the ratio of the energy needed at the wheels, E_{out}, to drive and brake a car of a given weight, M, on a specified test cycle to the total fuel energy, E_{in}, needed to drive the vehicle. Regenerative braking, if present, reduces the fuel needed to drive the car. Accessory power is not included in energy output. The TTW efficiency is calculated as $\eta_{TTW} = E_{out}/E_{in}$. For the vehicle configurations in Fig. 1, we keep E_{out} constant and calculate E_{in} by backward induction as
$$E_{in} = \frac{1}{\eta_{fp\eta e}}\left[\frac{E_{out}}{\eta_{dt}} - B\eta_{rb} + E_{ac} + E_{idle}\right]$$
where E_{idle} and E_{ac} are the energies required in the specified drive cycle for idling and for accessories, respectively. B is the recovered braking energy. The various efficiencies of different stages are η_{fp}, η_e, η_{dt}, and η_{rb} for fuel processing, engine or fuel cell, drive train, and regenerative braking, respectively. We focus on efficiency rather than the more common fuel economy because the efficiency is less sensitive to vehicle weight than fuel economy.
6. In 2001, General Motors (GM) collaborated with Argonne National Laboratories (ANL) to use ANL's Greenhouse Gases, Regulated Emissions, and Energy Use in Transportation (GREET) model. The report, "GM study: Well-to-wheel energy use and greenhouse gas emissions of advanced fuel/vehicle system, North American analysis," is referred to as the ANL/GM study and is available online at http://greet.anl.gov/publications.html.
7. M. A. Weiss, J. B. Heywood, E. M. Drake, A. Schafer, F. AuYeung, "On the road in 2020: A life cycle analysis of new automobile technologies," (MIT Energy Laboratory Report No. MIT EL 00-003, Cambridge, MA, 2000). We thank M. Weiss for helpful discussions about this work.
8. M. A. Weiss, J. B. Heywood, A. Schafer, V. K. Natarajan, "Comparative assessment of fuel cell cars" (MIT Laboratory for Energy and Environment Report No. 2003-001 RP, Cambridge, MA, 2003).
9. P. Ahlvi, A. Brandberg, Ecotraffic Research and Development, "Well to wheel efficiency for alternative fuels from natural gas to biomass," Vagverket (Swedish National Road Administration), Publication 2001: 85 (2001), appendix 1.8.
10. F. An, D. Santini, *SAE Tech. Pap*. 2003, no. 2003-01-0412 (2003).
11. B. Hohlein, G. Isenber, R. Edinger, T. Grube, *Handbook of Fuel Cells*, W. Vielstich, A. Gasteiger, A. Lamm, Ed. (Wiley, New York, 2003), vol. 3, chap. 21, p. 245.
12. More information is available at the DOE Web site: www.fueleconomy.gov/feg/atv.shtml.
13. We wish to keep the presentation of our model simple. The assumption of complete regenerative braking and reduction in idling losses is not realistic. However, improvement in ICE engine efficiency is also possible (*7*). The current performance of hybrid ICE passenger vehicles such as the Toyota Prius is impressive. Toyota reports TTW efficiency of the Prius as 32%, compared to 16% for a conventional ICE: www.toyota.co.jp/en/tech/environment/fchv/fchv12.html. Prius regenerative braking reportedly recaptures 30%; see www.ott.doe.gov/hev/regenerative.html.
14. For this case, there is no processor loss, and the FC stack efficiency improves to 55% because the FC functions better on pure hydrogen than reformate.
15. The NREL ADVISOR simulator is described online at www.ctts.nrel.gov/analysis. Use of the model is described in several publications at www.ctts.nrel.gov/analysis/reading_room.html; see, for example, (*20*, *21*).
16. GM quotes 15 to 20% fuel economy improvements in 2007 for hybrid Tahoe and Yukon sport utility vehicles. Not surprisingly, Toyota seems more optimistic about hybrids than GM.
17. In Europe, where fuel prices are much higher than in the United States, the advantage of hybrids over conventional ICEs is significantly greater.
18. The 2003 Energy Act, currently under consideration by Congress, would extend the time period for the hybrid car tax credit.
19. Quote taken from www.eere.energy.gov/vehiclesandfuels/.
20. T. Markel *et al*., *J. Power Sources* **110**, 225 (2002).
21. M. R. Cuddy, K. G. Wipke, *SAE Tech. Pap*. 1997, no. 970289 (1997).
22. This work was supported by the Alfred P. Sloan Foundation.

A Compound from Smoke That Promotes Seed Germination

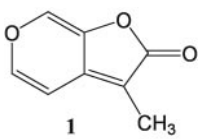
Gavin R. Flematti,[1]* Emilio L. Ghisalberti,[1] Kingsley W. Dixon,[2,3] Robert D. Trengove[4]

Smoke derived from burning plant material has been found to increase germination of a wide range of plant species from Australia, North America, and South Africa (*1*). We now report the identity of a compound, present in plant- and cellulose-derived smoke, that promotes germination of a variety of smoke-responsive taxa at a level similar to that of plant-derived smoke water.

The separation of the bioactive agent was facilitated by bioassay-guided fractionation with *Lactuca sativa* L. cv. Grand Rapids (*2*) and two smoke-responsive Australian species, *Conostylis aculeata* R. Br. (Haemodoraceae) and *Stylidium affine* Sonder. (Stylidiaceae) (*3*). Extensive fractionation of the relatively less complex, cellulose-derived smoke (from combustion of filter paper) resulted in the isolation of a compound that promotes seed germination (*4*). The structure of this compound was elucidated from mass spectrometry (MS) and spectroscopic data obtained by ^{1}H, ^{13}C, and two-dimensional (homonuclear correlation, heteronuclear single-quantum coherence, heteronuclear multi-bond correlation, and nuclear Overhauser effect spectroscopy) nuclear magnetic resonance (NMR) techniques. Confirmation of the structure as the butenolide 3-methyl-2*H*-furo[2,3-*c*]pyran-2-one (**1**) (Scheme 1) was achieved by synthesis. The presence of **1** in extracts of plant-derived smoke was confirmed by gas chromatography–MS analysis.

We compared the activity of the synthetic form of the butenolide (**1**) with that of plant-derived smoke water by testing it at a range of concentrations with the three bioassay species. The results (Fig. 1) show that **1** stimulated the germination of each test species to a level similar to that achieved with plant-derived smoke water. Furthermore, activity is demonstrated at very low concentrations (<1 ppb, 10^{-9} M). Testing of other smoke-responsive Australian species and smoke-responsive South African (e.g., *Syncarpha vestita*) and North American (e.g., *Emmenanthe penduliflora* and *Nicotiana attenuata*) species has further confirmed the activity of **1** (table S1).

The butenolide (**1**) conforms to the necessary ecological attributes of smoke that is produced from fires in natural environments. For example, the butenolide (**1**) is stable at high temperatures (its melting point is 118° to 119°C), water-soluble, active at a wide range of concentrations (1 ppm to 100 ppt), and capable of germinating a wide range of fire-following species. The butenolide is derived from the combustion of cellulose, which, as a component of all plants, represents a universal combustion substrate that would be present in natural fires.



Scheme 1.

Given the broad and emerging use of smoke as an ecological and restoration tool (*1*), the identification of **1** as a main contributor to the germination-promoting activity of smoke could provide benefits for horticulture, agriculture, mining, and disturbed-land restoration. In addition, the mode of action and mechanism by which **1** stimulates germination can now be investigated. In this context, it is useful to note that the natural product (+)-strigol, which promotes the germination of the parasitic weed *Striga* (*5*), is active at similar concentrations (10^{-9} M) and contains a butenolide moiety and additional conjugated functionality similar to those in **1**.

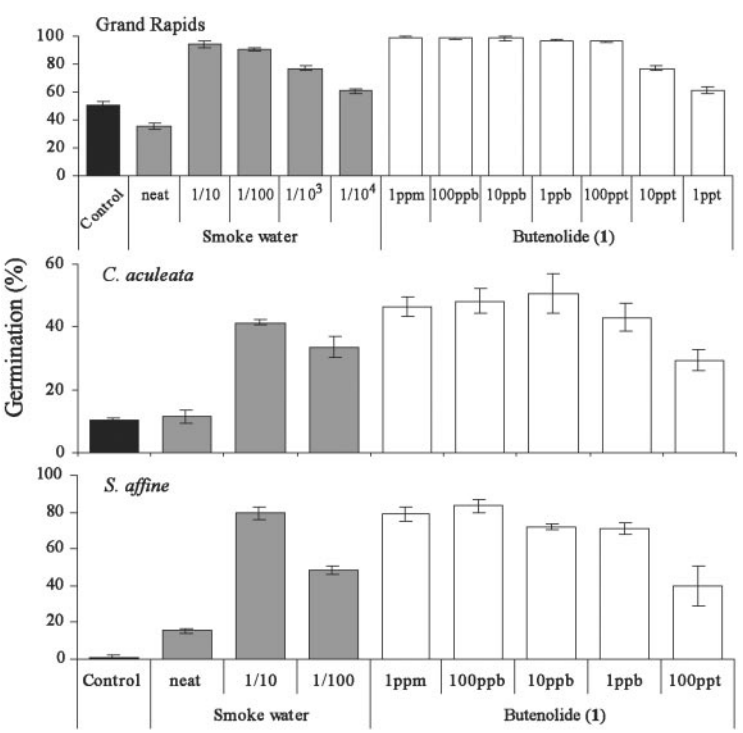


Fig. 1. Comparison of the germination response of plant-derived smoke water and butenolide (**1**) at different concentrations with three smoke-responsive species: Grand Rapids lettuce, *C. aculeata*, and *S. affine*. Water served as the control, and "neat" refers to undiluted smoke water. Values are means of three replicates ± SE.

References and Notes

1. N. A. C. Brown, J. Van Staden, *Plant Growth Regul.* **22**, 115 (1997).
2. F. E. Drewes, M. T. Smith, J. Van Staden, *Plant Growth Regul.* **16**, 205 (1995).
3. S. Roche, K. W. Dixon, J. S. Pate, *Aust. J. Bot.* **45**, 783 (1997).
4. Materials and methods are available as supporting material on *Science* Online.
5. S. C. M. Wigchert, B. Zwanenburg, *J. Agric. Food Chem.* **47**, 1320 (1999).

6. We thank L. T. Byrne for assistance with the structural elucidation of the active compound, D. Wege and S. K. Brayshaw for assistance with the synthetic approach, S. R. Turner and D. J. Merritt for assistance with germination trials, and Alcoa World Alumina and Iluka Resources for providing seeds of native species for testing.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1099944/DC1
Materials and Methods
Table S1

4 May 2004; accepted 25 June 2004
Published online 8 July 2004;
10.1126/science.1099944
Include this information when citing this paper.

[1]School of Biomedical and Chemical Sciences, [2]School of Plant Biology, The University of Western Australia, Crawley, WA 6009, Australia. [3]Kings Park and Botanic Garden, West Perth, WA 6005, Australia. [4]School of Engineering Science, Murdoch University, Rockingham, WA 6168, Australia.

*To whom correspondence should be addressed. E-mail: gflematt@chem.uwa.edu.au

Simulations of Jets Driven by Black Hole Rotation

Vladimir Semenov,[1] Sergey Dyadechkin,[1] Brian Punsly[2,3]*

The origin of jets emitted from black holes is not well understood; however, there are two possible energy sources: the accretion disk or the rotating black hole. Magnetohydrodynamic simulations show a well-defined jet that extracts energy from a black hole. If plasma near the black hole is threaded by large-scale magnetic flux, it will rotate with respect to asymptotic infinity, creating large magnetic stresses. These stresses are released as a relativistic jet at the expense of black hole rotational energy. The physics of the jet initiation in the simulations is described by the theory of black hole gravitohydromagnetics.

Most quasars radiate a small fraction of their emission in the radio band (radio quiet), yet about 10% launch powerful radio jets (a highly collimated beam of energy and particles) with kinetic luminosities rivaling or sometimes exceeding the luminosity of the quasar host (radio loud) (*1*). There is no clear theoretical understanding of the physics that occasionally switches on these powerful beams of energy in quasars. Powerful extragalactic radio sources tend to be associated with large elliptical galaxy hosts that harbor supermassive black holes [$\sim 10^9$ solar masses ($M_\odot$) (*2*). The synchrotron-emitting jets are highly magnetized and emanate from the environs of the central black hole, within the resolution limits of very long baseline interferometry (VLBI). Observations [see the supporting online material (SOM)] indicate that jets emitted from supermassive black hole magnetospheres may be required for any theory to be in accord with the data (*2–4*). Previous perfect magnetohydrodynamic (MHD) simulations of entire black hole magnetospheres have shown some suggestive results. For example, the simulations of (*5*) were the first to show energy extraction from the black hole, but there was no outflow of plasma. Numerical models of an entire magnetosphere involve a complex set of equations that reflect the interaction of the background spacetime with the plasma. In many instances, the only way to get any time evolution of the magnetosphere is to assume unphysical initial conditions (*5–8*). Even so, previous simulations have not shown a black hole radiating away its potential energy into a pair of bipolar jets (*5–10*). Most important, in previous efforts the underlying physics has been masked by the complexity of the simulation. Thus, this numerical work has done little to clarify the fundamental physics that couples the jet to the black hole.

We exploit the simplification that the full set of MHD equations in curved spacetime indicates that a magnetized plasma can be regarded as a fluid composed of nonlinear strings in which the strings are mathematically equivalent to thin magnetic flux tubes (*11*, *12*). In this treatment, a flux tube is thin by definition if the pressure variations across the flux tube are negligible compared to the total external pressure P (gas plus magnetic pressure), which represents the effects of the enveloping magnetized plasma (the magnetosphere). By concentrating the calculation on individual flux tubes in a magnetosphere, we can focus the computational effort on the physical mechanism of jet production (on all the field lines). Thus, we are able to elucidate the fundamental physics of black hole–driven jets without burying the results in the effort to find the P function. Our goal is to understand the first-order physics of jet production, not all of the dissipative second-order effects that modify the efficiency, and for these purposes the string depiction of MHD is suitable (see the methods section of the SOM). The physical mechanism of jet production is a theoretical process known as gravitohydromagnetics (GHM), in which the rotating spacetime geometry near the black hole drags plasma relative to the distant plasma in the same large-scale flux tube, thereby spring-loading the field lines with strong torsional stresses (*2*).

The frame-dragging potential of the rotating black hole geometry (described by Kerr spacetime) is responsible for driving the jets. The frame-dragging force is elucidated by the concept of the minimum angular velocity about the symmetry axis of the black hole as viewed from asymptotic infinity, $\Omega_p \geq \Omega_{min}$, where Ω_p is the angular velocity of the plasma [this frame is equivalent to Boyer-Lindquist (B-L) coordinates]. In flat spacetime, $\Omega_p > -c/r\sin\theta \equiv \Omega_{min} < 0$, where c is the speed of light and (r, θ) are spherical coordinates. By contrast, within the ergosphere (located between the event horizon at r_+ and the stationary limit, where r, θ, ϕ, and t are B-L coordinates), $\Omega_{min} > 0$. The ultimate manifestation of frame dragging is that all of the particle trajectories near the black hole corotate with the event horizon, $\Omega_{min} \rightarrow \Omega_H$ as $r \rightarrow r_+$ (the horizon boundary condition) (*2*). Now consider this Ω_{min} condition in a local physical frame at a fixed poloidal coordinate but rotating with $d\phi/dt \equiv \Omega$, so as to have zero angular momentum about the symmetry axis of the black hole, $m = 0$ (the ZAMO frames). In this frame (and in all physical frames; that is, those that move with a velocity less than c), a particle rotating at Ω_{min} appears to be rotating backward, azimuthally relative to black hole rotation at c, $\Omega_p = \Omega_{min} \Rightarrow c\beta^\phi = -c$. From equation 3.49 of (*2*), the mechanical energy of a particle in B-L coordinates (the astrophysical rest frame of the quasar), $\omega < 0$ if $\beta^\phi < -c(r^2 - 2Mr + a^2)^{1/2} \sin\theta/(\Omega\, g_{\phi\phi})$, in particular

$$\lim_{\beta^\phi \rightarrow -1} \omega = \mu_u^0 \frac{\Omega_{min}}{c} \sqrt{g_{\phi\phi}}\,,\ \Omega_{min} = \Omega - c(r^2 - 2Mr + a^2)^{1/2} \sin\theta/g_{\phi\phi} \quad (1)$$

In Eq. 1, the four-velocity of a particle or plasma in the ZAMO frames is given by $u^\lambda \equiv u^0(1, \beta)$, so u^0 is the Lorentz contraction factor that is familiar from special relativity; M is the mass of the black hole in geometrized units; a is the angular momentum per unit of mass of the black hole in geometrized units; μ is the specific enthalpy; and the B-L metric coefficient, $(g_{\phi\phi})^{1/2}$, is just the curved-space analog of the azimuthal measure, $r\sin\theta$, of flat spacetime up to a factor of a few. Because $\Omega_{min} > 0$ in the ergosphere, Eq. 1 implies that if a particle rotates so that $\Omega_p \approx \Omega_{min}$, then $\omega \ll 0$. Similarly, the specific angular momentum about the symmetry axis of the hole, $m = u^0\, \beta^\phi\, (g_{\phi\phi})^{1/2}$, is negative for these trajectories as well. Hence, the infall of these particles toward the black hole is tantamount to extracting its rotational energy.

Our simulation in Fig. 1 is of a perfect MHD (in terms of the field strength tensor $F^{\lambda\nu}\, u_\nu = 0$, $\forall\ \lambda$) plasma, threading an accreting poloidal magnetic flux tube that begins aligned parallel to the black hole spin axis. The perfect MHD condition means that the plasma can short out any electric fields that are generated in its rest frame. There are no existing numerical methods that can incorporate plasma dissipation into the black hole

[1]Institute of Physics, State University St. Petersburg, 198504 Russia. [2]Boeing Space and Intelligence Systems, Boeing Electron Dynamic Devices, 3100 West Lomita Boulevard, Post Office Box 2999, Torrance, CA 90509–1999, USA. [3]International Center for Relativistic Astrophysics (ICRA), University of Rome La Sapienza, I-00185 Roma, Italy.

*To whom correspondence should be addressed. E-mail: brian.m.punsly@boeing.com

magnetosphere (that is, that can go beyond perfect MHD); only the analytical work of (*2*) captures certain non-MHD effects, and their results agree with our simulations. In order to choose a realistic initial state, we note that ordered magnetic flux has been detected in the central 100 pc of some galaxies (*13*). Because Lens-Thirring torques concentrate the accreting plasma toward the plane orthogonal to the black hole spin axis, we expect the large-scale flux to become preferentially aligned with the spin axis (*14*). Long-term numerical simulations of magnetized accretion show a concentration of vertical flux near the black hole (*15*). Even an inefficient accretion of flux (reconnection of oppositely directed field lines and resistive diffusion) has been shown to lead to an enhanced flux distribution near the hole (*15*, *16*). Consequently, we have chosen our simulations to follow the accretion of a poloidal magnetic flux tube that begins aligned parallel to the black hole spin axis. The consistent magnetospheric field configuration is a result of the long-term accumulation of similar flux tubes. A rapidly spinning black hole is chosen in the simulation ($a/M = 0.9998$), as is often argued to be likely in a quasar (*17*).

Initially the flux tube is rotating with an angular velocity, Ω_F, equal to the local ZAMO angular velocity Ω_0: a rapidly decreasing function with radial coordinate (*2*). Thus, initially, $\Omega_0 \ll \Omega_H$, because the flux tube is far from the event horizon. The flux tube accretes toward the black hole under the influence of the gravitational force. The natural state of plasma motion (geodesic motion) induced by frame dragging is to spiral inward faster and faster as the plasma approaches corotation with the event horizon (the horizon boundary condition). By contrast, the natural state of plasma motion in a magnetic field is a helical Larmor orbit that is threaded onto the field lines. Generally, these two natural states of motion are in conflict near a black hole. These two strong opposing forces create a globally distributed torsion, creating the dynamical effect that drives the simulation (Fig. 1). The plasma far from the hole is still rotating slowly near Ω_0 in frame A. As the plasma penetrates the ergosphere, $\Omega_{min} \rightarrow \Omega_H$ as $r \rightarrow r_+$, and Ω_p must exceed Ω_0 in short order because $\Omega_0 \ll \Omega_H$ (within $t \sim 0.1\ GM/c^3$ after crossing the stationary limit). Thus, the ergospheric plasma gets dragged forward, azimuthally, relative to the distance portions of the flux tube, by the gravitational field. The back reaction of the field reestablishes the Larmor helical trajectories by torquing the plasma back onto the field lines with $J \times B$ forces (the cross-field current density J driven by this torsional stress is sunk within the enveloping magnetosphere). By Ampere's law, J creates a negative azimuthal magnetic field, B^Φ, upstream of the current flow. The $J \times B$ back-reaction forces eventually torque the plasma onto $\omega < 0$ trajectories as per Eq. 1. Frames B and C illustrate the fact that the B^Φ created in the ergosphere propagates upstream in the form of an MHD plasma wave at later times ($t > 70\ GM/c^3$) as more and more negative energy (the red portion of the field line) is created in the ergospheric region of the flux tube. In frame C, a bona fide jet (a collimated relativistic outflow of mass) emerges from the ergosphere.

The dynamo region for B^Φ in the ergosphere is expanded in frame D. Because $B^\Phi < 0$ upstream of the dynamo, there is an energy ($\sim -\Omega_F B^\Phi B^P$) and angular momentum ($\sim -B^\Phi B^P$) flux along B^P, away from the hole in the jet. The red portion of the field line indicates that the total plasma energy per particle is negative, $E < 0$ (because the magnetic field is primarily azimuthal near the black hole, $E \approx \omega + S/k$, where S is the poloidal Poynting flux along B, and k is the poloidal particle flux along B), downstream of the dynamo. Because $B^\Phi > 0$ downstream, the field transports energy and angular momentum toward the hole with the inflowing plasma. Thus, $S/k > 0$ in the downstream state, implying that $\omega \ll 0$ in order for $E < 0$, downstream. The $J \times B$ back-reaction forces on the twisted field lines torque the plasma onto trajectories with $\Omega_p \approx \Omega_{min}$. The ingoing $\omega \ll 0$ plasma extracts the rotational energy of the hole by Eq. 1; thus, black hole rotational inertia is powering the jet in the simulation. This is the fundamental physics of GHM (*2*).

Three additional movies of simulations are provided with different initial conditions [movie S3 has the same P as Fig. 1, but the field line geometry is different (the flux tube is highly inclined); movie S4 has a different pressure function, $P \sim (r - r_+)^{-2.2}$; and movie S5 has four flux tubes] to show the generality of the results of the simulation in movies S1 and S2 (Fig. 1).

It is important to make a connection between what is modeled here and what is observed in quasar jets. The jet can be considered as a bundle of thin flux tubes similar to those in our simulations (this is clearly visualized in movie S5). The jet is composed of Poynting flux and a relativistic outflow of particles. In the simulation of Fig. 1, jet plasma attains a bulk flow Lorentz factor of 2.5, at $r \sim 50\ GM/c^2$ (which is $\sim 10^{16}$ cm from a $10^9\ M_\odot$ black hole, which is assumed in all of the following estimates) in frame C (at $t \sim 250\ GM/c^3$). In quasars, VLBI observations indicate that jet Lorentz factors are in the range of 2 to 30, with most ~10 (*18*). The resolution of VLBI is on the order of 100 to 1000 times the length of the jet in the simulations. Physically, it is believed that Poynting flux is required to accelerate the jet plasma to very high bulk flow Lorentz factors of ~ 10 on parsec scales (*19*, *20*). Furthermore, the magnetic tower created by B^Φ in Fig. 1, frame C, in combination with the

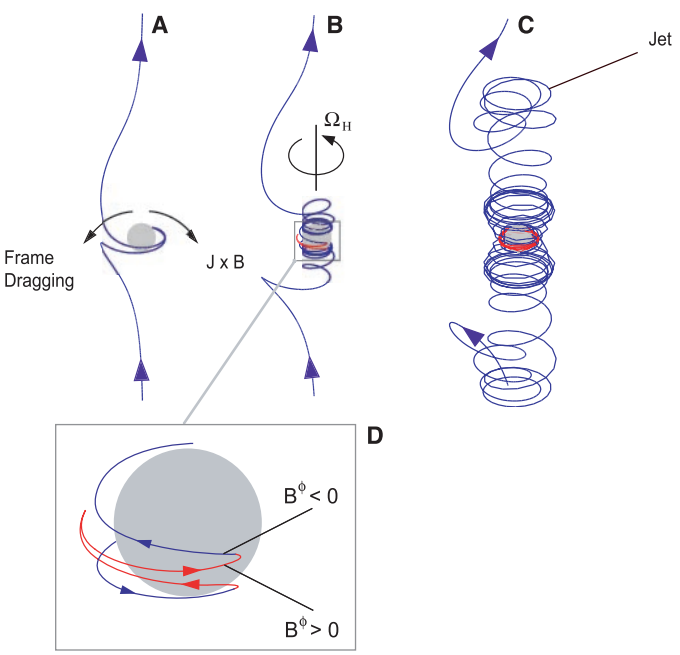


Fig. 1. Black hole–driven jet. A jet is produced on the magnetic flux tubes that experience the torsional stress induced by the opposition of the gravitational frame-dragging force with the $J \times B$ electromagnetic force, in which J is the current density and B is the magnetic field. The simulation is of a single flux tube in an enveloping magnetosphere of similar flux tubes. The entire flux tube rotates in the same sense as the black hole, where Ω_H is the angular velocity of the event horizon as viewed from asymptotic infinity. The red portions of the field line indicate plasma with negative total energy, as viewed from asymptotic infinity. The black hole has a radius $r \approx GM/c^2$, which is $\approx 1.5 \times 10$ cm (*14*) for a 10^9-$M_\odot$ black hole. Frame **A** is a snapshot of the initiation of negative energy generation in the ergosphere: the outer boundary of the ergosphere is given by $r \approx GM/c^3\ (1 + \sin\theta)$. This effect is seen in (*5*), but that simulation ends at this early stage. In frame **B**, an outgoing Poynting flux emerges from the ergosphere, and frame **C** shows a well-formed jet. The time lapses between frames, as measured by a distant stationary observer, are: A to B, $t = 78\ GM/c^3$ and A to C, $t = 241\ GM/c^3 \approx 13$ days. The simulation ends with a pair of jets, each over $50\ GM/c^2$ in length. Frame **D** is a close-up of the dynamo region, in which B^Φ is created in the jet (that is, where B^Φ changes sign). The pure Alfven speed [a measure of the ratio of magnetic energy to plasma inertia, $U_A = B^P/(4\pi n \mu c^2)^{1/2}$] in the simulation is $U_A \approx 12c$ to $13c$ in the region in which Poynting flux is injected into the jet just above the dynamo.

poloidal field component B^P, naturally provides stable hoop stresses that are the only known collimation mechanism for the large-scale jet morphology of quasars (*21*, *22*). The nonrelativistic outer layer observed in some jets can be supported by a coexisting, enveloping, relatively low-power wind or jet from the accretion disk (*23*).

The jets composed of a bundle of strings like those in the simulation are energetic enough to power quasar jets. The Poynting flux transported by a pair of bipolar collimated jets is approximately (*2*)

$$S \approx \frac{[\Omega_F \Phi]^2}{2\pi^2 c} \quad (2)$$

The total magnetic flux in the jet is Φ. The field line angular velocity Ω_F varies from near zero at the outer boundary of the ergosphere to the horizon angular velocity $\Omega_H \approx 10^{-4}$ s^{-1} in the inner ergosphere. Thus, $S \sim (\Omega_H \Phi)^2 = (a\Phi/2Mr_+)^2 \sim (a\Phi/M)^2$ for rapidly rotating black holes. Consequently, there are three important parameters that determine jet power: M, a, and B. For rapidly spinning black holes, the surface area of the equatorial plane in the ergosphere becomes quite large; for $a/M = 0.996$, it is $\approx 30\ M^2$ (*2*). Various accretion flow models yield a range of achievable ergospheric field strengths $B \sim 10^3$ G to 2×10^4 G that equate to $\Phi \sim 10^{33}$ G-cm^2 to 10^{34} G-cm^2 (*2*, *14*, *24*). Inserting these results into Eq. 2 yields a jet luminosity of $\approx 10^{45}$ to 5×10^{47} ergs/s. This is consistent with the estimates of the kinetic luminosity of powerful quasar jets (*1*). The maximum value of the flux noted above occurs when the persistent accretion of magnetic flux has a pressure that is capable of pushing the inner edge of the accretion disk out of the ergosphere (known as magnetically arrested accretion) (*15*, *16*). This maximum flux inserted in Eq. 2 equates to a jet power that is $\approx$ 5 to 25 times the bolometric thermal luminosity of the accretion flow in the disk models considered in (*2*, *24*).

If 10% of the central black holes in quasars were magnetized by the accretion of vertical flux, this would explain the radio loud/radio quiet quasar dichotomy. To elevate this above a conjecture requires observational corroboration of the putative magnetosphere in radio loud quasars. A significant flux trapped between the black hole and the accretion disk should modify the innermost regions of the accretion flow. Thus, one can look for a distinction between radio loud and radio quiet quasar thermal emission at the highest frequencies. There might already be evidence to support this. The accretion flow radiation has a high-frequency tail in the EUV (extreme ultraviolet). Hubble Space Telescope (HST) data indicate that radio quiet quasars have an EUV excess as compared to radio loud quasars (*25*). The EUV suppression has been explained by the interaction of the magnetic field with the inner edge of the disk, displacing the EUV-emitting gas in radio loud quasars (*26*).

The simulations presented here explain five important observations of radio loud quasars: the production of a collimated jet (based on radio observations); a power source (the black hole) that is decoupled from the accretion flow properties to first order [broadband radio-to-ultraviolet observations indicate that a quasar can emit most of its energy in a jet without disrupting the radiative signatures of the accretion flow (see the SOM)]; the suppression of the EUV in radio loud quasars (from HST observations); the relativistic velocity of the jet (from VLBI data); and the maximal kinetic luminosity of the quasar jets (from broadband radio and x-ray observations of radio lobes). The GHM process might also drive jets in other systems such as microquasars or gamma-ray bursts (*27*, *28*). However, microquasars show correlations between accretion disk emission and jet properties, unlike quasars (*29*). Consequently, there is no strong observational reason to prefer the black hole over the accretion disk as the primary power source for microquasars, as there is with quasar jets.

References

1. K. Blundell, S. Rawlings, *Astron. J.* **119**, 1111 (2000).
2. B. Punsly, *Black Hole Gravitohydromagnetics* (Springer-Verlag, New York, 2001).
3. M. J. Rees, E. S. Phinney, M. C. Begelman, R. D. Blandford, *Nature* **295**, 17 (1982).
4. M. C. Begelman, R. D. Blandford, M. J. Rees, *Rev. Mod. Phys.* **56**, 255 (1984).
5. S. Koide, K. Shibata, T. Kudoh, D. L. Meier, *Science* **295**, 1688 (2002).
6. S. Koide, *Phys. Rev. D* **67**, 104010 (2003).
7. S. Koide, *Astrophys. J. Lett.* **606**, L45 (2004).
8. S. Komissarov, *Mon. Not. R. Astron. Soc.* **350**, 1431 (2004).
9. S. Hirose, J. Krolik, J. De Villiers, *Astrophys. J.* **606**, 1083 (2004).
10. J. McKinney, C. Gammie, *Astrophys. J.*, in press; preprint available at http://arxiv.org/abs/astro-ph/0404512.
11. M. Christensson, M. Hindmarsh, *Phys. Rev.* **D60**, 063001-1 (1999).
12. V. S. Semenov, S. A. Dyadechkin, I. B. Ivanov, H. K. Biernat, *Phys. Scripta* **65**, 13 (2002).
13. T. J. Jones, *Astron. J.* **120**, 2920-2927 (2000).
14. D. Macdonald, K. Thorne, R. Price, X.-H. Zhang, in *Black Holes The Membrane Paradigm*, K. Thorne, R. Price, D. Macdonald, Eds. (Yale Univ. Press, New Haven, CT, 1986), pp. 121–145.
15. I. V. Igumenshchev, R. Narayan, M. A. Abramowicz, *Astrophys. J.* **592**, 1042 (2003).
16. R. Narayan, I. V. Igumenshchev, M. A. Abramowicz, *Publ. Astron. Soc. Jpn.* **55**, 69 (2003).
17. J. Bardeen, *Nature* **226**, 64 (1970).
18. K. Kellerman *et al. Astrophys. J.* **609**, 539 (2004).
19. R. V. E. Lovelace, M. M. Romanova, *Astrophys. J. Lett.* **596**, 159 (2003).
20. V. Nektarios, A. Konigl, *Astrophys. J.* **605**, 656 (2004).
21. G. Benford, *Mon. Not. R. Astron. Soc.* **183**, 29 (1978).
22. P. Hardee, in *Cygnus A–Study of a Radio Galaxy*, C. L. Carilli, D. E. Harris, Eds. (Cambridge Univ. Press, New York, 1996), pp. 113–120.
23. G. Henri, G. Pelletier, *Astrophys. J.* **383**, L7 (1991).
24. F. Casse, R. Keppens, *Astrophys. J.* **601**, 90 (2004).
25. W. Zheng, G. A. Kriss, R. C. Telfer, J. P. Grimes, A. F. Davidsen, *Astrophys. J.* **475**, 469 (1997).
26. B. Punsly, *Astrophys. J.* **527**, 609 (1999).
27. S. Eikenberry *et al.*, *Astrophys. J.* **494**, L61 (1998).
28. J. Katz, *Astrophys. J.* **490**, 633 (1997).
29. R. Fender, in *Compact Stellar X-Ray Sources*, W. H. G. Lewin, M. van der Klis, Eds. (Cambridge Univ. Press, Cambridge, in press) (preprint available at http://arxiv.org/abs/astro-ph/0303339).

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/978/DC1
Methods
SOM Text
Figs. S1 and S2
References
Movies S1 to S5

24 May 2004; accepted 8 July 2004

Localization of Fractionally Charged Quasi-Particles

Jens Martin,[1]* Shahal Ilani,[1]† Basile Verdene,[1] Jurgen Smet,[2] Vladimir Umansky,[1] Diana Mahalu,[1] Dieter Schuh,[3] Gerhard Abstreiter,[3] Amir Yacoby[1]

An outstanding question pertaining to the microscopic properties of the fractional quantum Hall effect is understanding the nature of the particles that participate in the localization but that do not contribute to electronic transport. By using a scanning single electron transistor, we imaged the individual localized states in the fractional quantum Hall regime and determined the charge of the localizing particles. Highlighting the symmetry between filling factors 1/3 and 2/3, our measurements show that quasi-particles with fractional charge $e^* = e/3$ localize in space to submicrometer dimensions, where e is the electron charge.

The quantum Hall effect (QHE) arises when electrons confined to two dimensions are subject to a strong perpendicular magnetic field. The magnetic field quantizes the kinetic energy and leads to the formation of Landau levels (LL). Energy gaps appear in the spectrum whenever an integer number of LLs is filled. Disorder broadens the LLs and gives rise to bands of extended states surrounded by bands of localized states. Localization plays a fundamental role in the universality and robustness of quantum Hall phenomena. In the localized regime, as the den-

sity of electrons increases, only localized states are being populated and hence the transport coefficients remain universally quantized (*1*). Conversely, when the Fermi energy lies within the bands of extended states, the transport coefficients vary indicating transitions between quantum Hall phases (*2*, *3*).

When the filling of the lowest LL is less than one, the fractional quantum Hall effect (FQHE) emerges (*4*) as a result of Coulomb interactions between electrons. The Coulomb interactions give rise to a new set of energy gaps occurring at fractional fillings $\nu = \frac{q}{2q \pm 1}$ with q being an integer. Soon after the discovery of the FQHE, a theory was presented for the ground state properties of these unique phases (*5*), in which the low energy excitations above these ground states are fractionally charged with $e^* = \frac{e}{2q \pm 1}$. Only recently, through use of resonant tunneling and shot noise measurements, were such fractionally charged quasi-particles shown to exist experimentally (*6*–*10*). These measurements convincingly demonstrate that the fractionally charged quasi-particles participate in transport and are hence extended along the edge of the sample. However, what is the nature of the particles that participate in the localization and do not contribute to transport? We address this question with the use of a scanning single electron transistor (SET), which enables us to detect directly the position and charge of the localizing particles across the various integer and fractional quantum Hall phases.

Our experimental method is described in detail elsewhere (*11*, *12*). A SET is used to measure the local electrostatic potential of a two-dimensional electron gas (2DEG) (*13*). At equilibrium, changes in the local electrostatic potential result from changes in the local chemical potential, which can be induced, for example, by varying the average electron density (*14*) with the use of a back gate. The local derivative of the chemical potential, μ, with respect to the electronic density, $d\mu/dn$, is inversely proportional to the local compressibility and depends strongly on the nature of the underlying electronic states (*15*, *16*). In the case of extended electrons, charge is spread over large areas; hence, the chemical potential will follow continuously the variations in density, and the measured inverse compressibility will be small and smooth. However, in the case of localized electrons the local charge density can increase only in steps of e/ξ^2, where ξ is the localization length. Hence, the system becomes compressible only when a localized state is being populated, producing a jump in the local chemical potential and a spike in its derivative, the inverse compressibility. Thus, by using the scanning SET in the localized regime, we were able to image the position and the average density at which each localized state is populated. Three different GaAs-based 2DEGs with peak mobilities of 2×10^6 cm^2/ V^{-1} s^{-1} (V6-94 and V6-131) and 8×10^6 cm^2/ V^{-1} s^{-1} (S11-27-01.1) were studied.

We start by reviewing the properties of localized states in the integer QHE. Figure 1A shows a typical scan of $d\mu/dn$ as function of the average electron density, n, and position, x, near the integer quantum Hall phase $\nu = 1$. Each black arc corresponds to the charging line of an individual localized state. At any particular location, as the density is varied through $\nu = 1$, electrons occupying localized states give rise to negative spikes. The tip detects only charging of localized states situated directly underneath it. The measured spatial extent of each localized state (black arc) is therefore determined by the size of the localized state convoluted with the spatial resolution of the tip. A small tip bias is responsible for the arcing shape of each charging line. Figure 1A shows that electrons do not localize randomly in space but rather pile up at particular locations. Moreover, the spacing within each charging spectrum is regular. Such charging spectra are reminiscent of Coulomb blockade physics, where charge quantization governs the addition spectrum of a quantum dot. In the microscopic description of localization in the integer QHE (*17*), the formation of dots was explained with use of a simple model that incorporates Coulomb interaction between electrons and thereby accounts for changes in the screening properties of the 2DEG as the filling factor is varied (*18*–*21*). For completeness, we briefly sketch the model for $\nu = 1$ because it will prove to be essential for understanding the measured spectra at fractional filling factors.

An intuitive picture of localization driven by Coulomb interaction is obtained by tracing the self-consistent density distribution as a function of B and n. Far from integer filling, the large compressibility within an LL provides nearly perfect screening of the disorder potential. This is accomplished by creating a nonuniform density profile with a typical length scale larger than the magnetic length, $l_m = \sqrt{h/(eB)}$ (Fig. 1B). The corresponding potential distribution in the plane of the 2DEG is equal and opposite in sign to the disorder potential. Because of a large energy gap between the LLs, the density, n_{LL}, cannot exceed one electron per flux quanta, $n_{max} = B/\phi_0$, and is therefore constrained by $0 \leq n_{LL} \leq n_{max}$. At the center of the LL (far from integer filling), this constraint is irrelevant because the variations in density required for screening are smaller than n_{max}, thus leading to nearly perfect screening (Fig. 1B). Each electron added to the system experiences this flat potential and thus, within this approximation, is completely delocalized. With increasing filling factor, the density distribution increases uniformly, maintaining its spatial structure. However, near $\nu = 1$ the average density approaches n_{max} and the required density distribution for perfect screening exceeds n_{max} at certain locations. Because of a large energy gap between the LLs, n_{LL} cannot exceed n_{max}, and the density at these locations becomes pinned to n_{max}. Local incompressible regions are formed in which the bare disorder potential is no longer screened, coexisting with compressible regions where the LL is still only partially full [$0 < n_{LL}(x) < n_{max}$, where

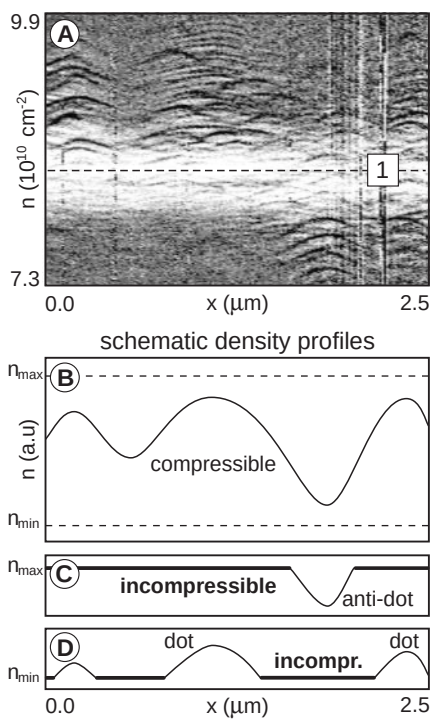


Fig. 1. (**A**) Density-position scan of $d\mu/dn$ near $\nu = 1$ [$B = 3.5$ T, temperature (T) = 300 mK, sample V6-94]. Each black line corresponds to a localized state. The central white area corresponds to the incompressible region near complete filling of the first LL (the dashed line marks $\nu = 1$). Localized states group into families, which appear as QD spectra above (and antidot spectra below) the incompressible region. (**B**) Far from integer filling, the large compressibility within an LL provides nearly perfect screening at the cost of a nonuniform density profile. Here, a schematic density profile is shown. n_{max} and n_{min} indicate the maximum and minimal densities, respectively, within an LL. a.u., arbitrary units. (**C**) In case the average density is slightly less than complete filling, the 2DEG becomes incompressible whenever the maximum allowed density n_{max} is reached and remains compressible only in a small area. This area forms an antidot whose energy levels are determined by the charging energy. (**D**) Once the average density is slightly higher than complete filling, compressible areas appear in the next LL and form QDs.

[1]Weizmann Institute of Science, Condensed Matter Physics, 76100 Rehovot, Israel. [2]Max-Planck Institut für Festkörperforschung, D-70569 Stuttgart, Germany. [3]Walter-Schottky Institut, Technische Universität München, D-85748 Garching, Germany.

*To whom correspondence should be addressed. E-mail: jens.martin@weizmann.ac.il

†Present address: Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA.

x is the position] and the local potential is screened (Fig. 1C). Once an incompressible region surrounds a compressible region, it behaves as an antidot whose charging is governed by Coulomb blockade physics. Antidots are therefore formed in the regions of lowest local density, and as the filling increases further the antidot will be completely emptied. Spatially separated from such local density minima are the maxima of the density profile, where the filling of the next LL will first occur, now as quantum dots (QDs) (Fig. 1D).

An important consequence of the above model is that the spectra of localized states are defined only by the bare disorder potential, the presence of an energy gap, causing an incompressible quantum Hall liquid surrounding the QD, and the charge of the localizing particles. It is the character of the boundary that determines the addition spectra of the quantum dot, and it is because of Coulomb blockade that the charge directly determines the number of charging lines and their separation and their strength. Therefore, the number of localized states does not depend on the magnetic field. A different magnetic field will simply shift n_{max} and hence the average density at which localization commences. Therefore, the local charging spectra will evolve as a function of density and magnetic field exactly parallel to the quantized slope of the quantum Hall phases, $dn/dB = \nu e/h$ (Fig. 2, A, I and III).

The presence of an energy gap at fractional filling factors implies that the model for localization described above may also apply to the fractional quantum Hall regime. As in the integer case, the local charging spectra contain a fixed number of localized states (independent of B) that evolve in density and magnetic field according to $dn/dB = \nu e/h$. The difference between the integer and fractional case is that now the slopes are quantized to fractional values (Fig. 2, A, II and IV, and B). The invariance of the local charging spectra along fractional filling factors can also be inferred from Fig. 2, C and D, where we show that the spatially dependent charging spectra at $\nu = 1/3$ for two different magnetic fields are identical. Our observations confirm that the microscopic mechanism for localization in the fractional quantum Hall regime is identical to that of the integer case; i.e., localization is driven by Coulomb interaction. The QDs that appear near integer or fractional filling factors are, therefore, identical. Their position, shape, and size are solely determined by the underlying bare disorder potential. This conclusion allows us to determine the charge of localizing particles in the fractional quantum Hall regime by comparing the charging spectra of integer and fractional filling factors. Charging spectra corresponding to the localization of electrons should look identical to those seen at integer filling. On the other hand, the charging spectra of fractionally charged quasi-particles, with $e^* = e/3$ for example, would have a denser spectrum with three times more charging lines, i.e, localized quasi-particle states. We emphasize that the slope of the localized states in the n-B plane merely indicates the filling factor they belong to rather than their charge.

The spatially resolved charging spectra for integer $\nu = 1$ and $\nu = 3$ and fractional $\nu = 1/3$ and $\nu = 2/3$ are shown in Fig. 3. All the scans are taken along the same line in space and cover the same density interval. The scans differ only in the starting density and the applied magnetic field. The spatially resolved charging spectra for integer fillings in Fig. 3, A and B, look identical. As expected, despite the different filling factors the measured spectra of localized states appear at the same position in space and with identical spacing. Figure 3, C and D, shows spatially resolved charging spectra at $\nu = 1/3$ and $\nu = 2/3$. Here also, the two spectra look identical. Moreover, the charging spectra are seen at the same locations as in the integer case. The only difference between the spectra taken at integer and fractional filling is the number of charging lines within a given range of densities. At $\nu = 1/3$ and $\nu = 2/3$, there are three times more charging lines and the separation between charging lines is three times smaller. This can be also seen in Fig. 3, E to H.

Our model assumes large energy gaps relative to the bare disorder potential. In practice, however, the gap for fractional filling factors is considerably smaller than in the integer case. In order to have a comparable gap in the integer and fractional regime, we first measured the integer quantum Hall phases at low magnetic field and then measured the fractions at high field. Such a comparison is meaningful because the number of localized states is independent of magnetic field. In order to eliminate any doubt that the higher number of localized states in the fractional regime result from the measurement at higher fields, we also compare in Fig. 3, G and H, spectra of $\nu = 1$ and $\nu = 1/3$ measured at the same magnetic field, $B = 7.25$ T. Regardless of field or density, the spectrum of $\nu = 1/3$ is three times denser than that of $\nu = 1$.

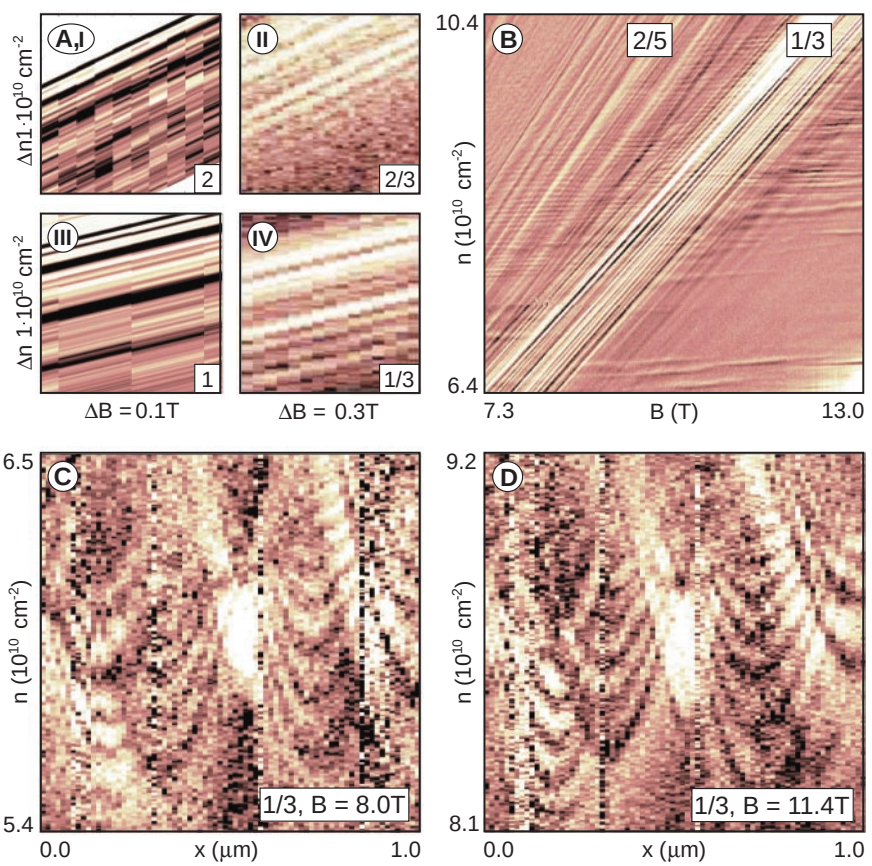


Fig. 2. (**A**) Density-magnetic field scans for integer $\nu = 2$ (I) and $\nu = 1$ (III) and for fractional $\nu = 2/3$ (II) and $\nu = 1/3$ (IV) (sample V6-131). All scans cover the same density range Δn. The range in magnetic field is three times larger for the fractional fillings. Each black line corresponds to a localized state. (**B**) This density-magnetic field scan covers a much larger range in field and density than (A) and shows localized states for $\nu = 1/3$ and $\nu = 2/5$ (sample S11-27-01.1, $T < 100$ mK). Even for a change of the magnetic field by a factor of 2, the number of localized states does not change. (**C**) Density-position scan for $\nu = 1/3$ at $B = 8$ T (sample V6-131, $T = 300$ mK). Just as for the integer QHE, the localized states appear as QD spectra. (**D**) Scan at the same position as in (C) but at $B = 11.4$ T. The same QD spectra are recovered at higher density. They are better resolved as a result of the enhanced gap size at higher fields.

Our results constitute direct evidence that quasi-particles with charge $e/3$ localize at $\nu = 1/3$ and $\nu = 2/3$. Moreover, our results highlight the symmetry between filling factors 1/3 and 2/3, indicating directly that at $\nu = 2/3$ the quasi-particle charge is $e/3$. In contrast to the experiments on resonant tunneling across an artificial antidot (*6*, *7*), our measurements probe generic localized states in the bulk of the 2DEG. The localization area, ξ^2, can be inferred from the charging spectra by using the condition for charge neutrality: $e^* = e\Delta n\xi^2$, where is the measured spacing between charging lines in units of density. For the left spectrum in Fig. 1D, $e^* = e/3$ and $\Delta n \approx 1 \times 10^{11}$ cm^{-2}, which corresponds to $\xi \approx 200$ nm, indicating quasi-particle localization to submicrometer dimensions. The extracted localization length provides further support for the validity of our model, which assumes long-range disorder relative to the magnetic length.

So far we have concentrated on the level spacing of the charging spectra. We now turn to address the amplitude of a single charging event. The integrated signal across a single charging event (spike) is the change in the local chemical potential associated with the addition of a single charge quantum, given by $\Delta\mu = e^*/C_{tot}$, where C_{tot} is the total capacitance of the QD. Because the capacitance between the QD and its surrounding is unknown experimentally, one cannot use this jump in chemical potential to determine the absolute charge of the localized particle. However, knowing that the QDs formed at integer and fractional filling are identical, one expects C_{tot} to be unchanged. Therefore, the ratio of $\Delta\mu$ at different filling factors is a measure of the ration of quasi-particle charge. Figure 4 shows a cross section of the measured $d\mu/dn$ through the center of a QD and the integrated signal $\Delta\mu$. One can clearly see that the step height in the fractional regime is only about 1/3 of the step height in the integer regime, confirming the localization of quasi-particles with $e^* = e/3$ for both $\nu = 1/3$ and $\nu = 2/3$.

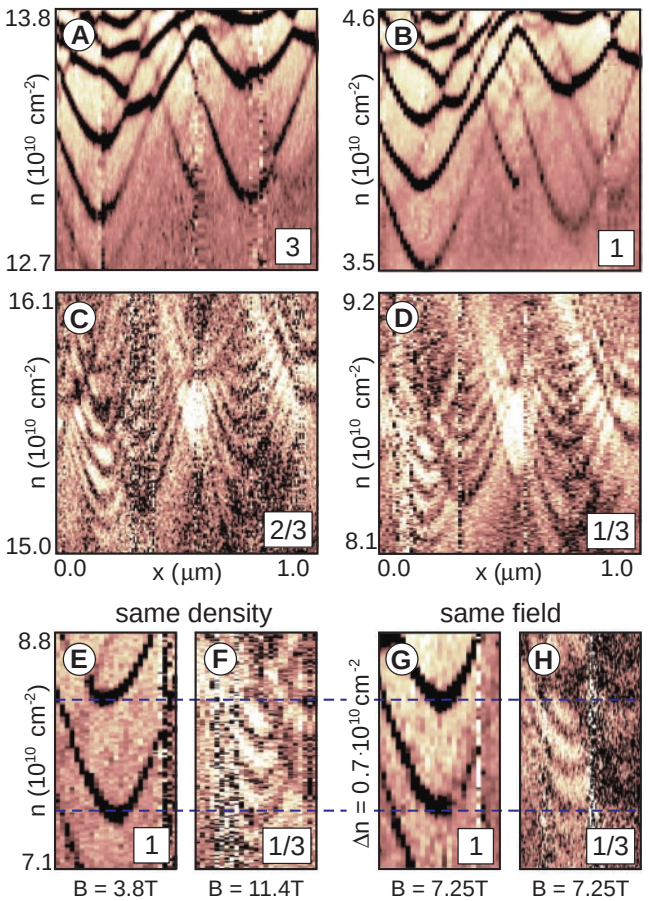


Fig. 3. (**A**) Density-position scan for $\nu = 3$ ($B = 1.9$ T, $T = 300$ mK, sample V6-131). The level spacing decreases slightly with increasing density, indicating the spectra of QDs. (**B**) Density-position scan at the same position as (A) for $\nu = 1$ ($B = 1.9$ T). (**C**) Density-position scan for $\nu = 2/3$ ($B = 10$ T) at the same position as the integer scans. (**D**) Density-position scan for $\nu = 1/3$ ($B = 11.4$ T, same as in Fig. 2D). Comparison between integer and fractional filling factors reveals three times denser LS spectra for the fractional regime. (**E**) Left dot at $\nu = 1$ measured at the same density (one-third the magnetic field) as scan for $\nu = 1/3$ in (**F**). (**G**) Left dot at $\nu = 1$ measured at the same magnetic field (3 times the density) as the scan for $\nu = 1/3$ in (**H**). The dashed lines serve as a guide to the eye to emphasize the difference in the level spacing between $\nu = 1$ and $\nu = 1/3$. The level spacing is independent of magnetic field and density.

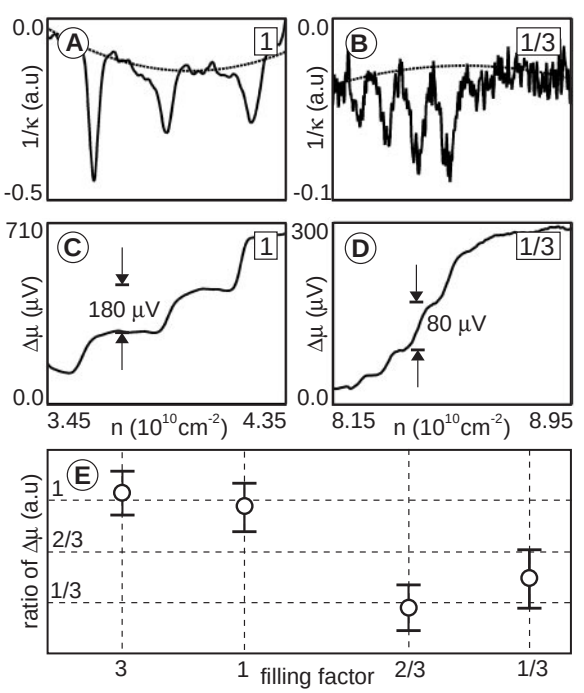


Fig. 4. (**A**) Cross section of the measured compressibility, κ, through the center of the left dot for $\nu = 1$ (Fig. 3B). For integration, the background compressibility (dotted line) is subtracted. (**B**) Cross section through the center of the left dot for $\nu = 1/3$ (Fig. 3D). (**C**) Integrated signal (chemical potential μ) with background correction for $\nu = 1$. Each step corresponds to one electron or quasi-particle entering the QD. (**D**) Integrated signal with background correction for $\nu = 1/3$. (**E**) Step height for the left dot for the various integer and fractional filling factors.

References and Notes

1. R. E. Prange, S. M. Girvin, Eds., *The Quantum Hall Effect* (Springer-Verlag, New York, ed. 2, 1990), chap. 1.
2. S. Kivelson, D. H. Lee, S. C. Zhang, *Phys. Rev. B* **46**, 2223 (1992).
3. B. Huckestein, *Rev. Mod. Phys.* **67**, 357 (1995).
4. D. C. Tsui, H. L. Stormer, A. C. Godard, *Phys. Rev. Lett.* **48**, 1559 (1982).
5. R. B. Laughlin, *Phys. Rev. Lett.* **50**, 1359 (1983).
6. V. J. Goldman, B. Su, *Science* **267**, 1010 (1995).
7. J. D. F. Franklin *et al.*, *Surf. Sci.* **361-362**, 17 (1996).
8. L. Saminadayar, R. V. Glattli, Y. Jin, B. Etienne, *Phys. Rev. Lett.* **79**, 2526 (1997).
9. M. Reznikov, R. de Picciotto, T. G. Griffiths, M. Heiblum, V. Umansky, *Nature* **399**, 238 (1999).
10. R. de Picciotto *et al.*, *Nature* **389**, 162 (1997).
11. M. J. Yoo *et al.*, *Science* **276**, 579 (1997).
12. A. Yacoby, H. F. Hess, T. A. Fulton, L. N. Pfeiffer, K. W. West, *Solid State Commun.* **111**, 1 (1999).
13. N. B. Zhitenev *et al.*, *Nature* **404**, 473 (2000).
14. J. P. Eisenstein, L. N. Pfeiffer, K. W. West, *Phys. Rev. B* **50**, 1760 (1994).
15. S. Ilani, A. Yacoby, D. Mahalu, H. Shtrikman, *Phys. Rev. Lett.* **84**, 3133 (2000).
16. S. Ilani, A. Yacoby, D. Mahalu, H. Shtrikman, *Science* **292**, 1354 (2001).
17. S. Ilani *et al.*, *Nature* **427**, 328 (2004).
18. A. L. Efros, A. F. Ioffe, *Solid State Commun.* **67**, 1019 (1988).
19. D. B. Chklovskii, P. A. Lee, *Phys. Rev. B* **48**, 18060 (1993).
20. N. R. Cooper, J. T. Chalker, *Phys. Rev. B* **48**, 4530 (1993).
21. I. Ruzin, N. Cooper, B. Halperin, *Phys. Rev. B* **53**, 1558 (1996).

22. This work is supported by the Israel Science Foundation, the Minerva Foundation, and the Fritz Thyssen Stiftung.

5 May 2004; accepted 9 July 2004

DNA-Functionalized Nanotube Membranes with Single-Base Mismatch Selectivity

Punit Kohli, C. Chad Harrell, Zehui Cao, Rahela Gasparac, Weihong Tan, Charles R. Martin*

We describe synthetic membranes in which the molecular recognition chemistry used to accomplish selective permeation is DNA hybridization. These membranes contain template-synthesized gold nanotubes with inside diameters of 12 nanometers, and a "transporter" DNA-hairpin molecule is attached to the inside walls of these nanotubes. These DNA-functionalized nanotube membranes selectively recognize and transport the DNA strand that is complementary to the transporter strand, relative to DNA strands that are not complementary to the transporter. Under optimal conditions, single-base mismatch transport selectivity can be obtained.

In both biology and technology, molecular recognition (MR) chemistry is used to selectively transport chemical species across membranes. For example, the transmembrane proteins that transport molecules and ions selectively across cell membranes contain MR sites that are responsible for this selective-permeation function (*1*–*3*). In a similar manner, MR agents such as antibodies have been incorporated into synthetic membranes so that the membranes will selectively transport the species that binds to the MR agent (*4*, *5*). However, there appear to be no previous examples of either biological or synthetic membranes where nucleic acid hybridization is used as the MR event to facilitate DNA or RNA transport through the membrane (*6*, *7*). If such membranes could be developed, they might prove useful for DNA separations and for the sensors needed, for example, in genomic research.

We describe synthetic MR membranes for selective DNA transport. We prepared these membranes by incorporating a "transporter" DNA, in this case a DNA-hairpin (*8*, *9*) molecule (Table 1), within the nanotubes of a gold nanotube membrane (*10*, *11*). We found that these membranes selectively recognized and transported the DNA molecule that was complementary to the transporter DNA. The rate of transport (flux) of the complementary strand was higher than the fluxes of permeating DNA molecules (Table 1) that contained as few as a single-base mismatch with the transporter DNA.

We prepared gold nanotube membranes with the template synthesis method (*12*), by electrolessly depositing gold along the pore walls of a polycarbonate template membrane (*10*, *11*). The template was a commercially available filter (Osmonics), 6 μm thick, with cylindrical, 30-nm-diameter pores and 6×10^8 pores per cm^2 of membrane surface area. The inside diameters of the gold nanotubes deposited within the pores of the template can be controlled by varying the deposition time. The membranes used here contained gold nanotubes with inside diameters of 12 ± 2 nm, as determined by a gas-flux measurement on three identical samples (*10*).

We chose a DNA hairpin (*8*, *9*) as our transporter strand. DNA hairpins contain a complementary base sequence at each end of the molecule (Table 1), and in an appropriate electrolyte solution, intramolecular hybridization causes a closed stem/loop structure to form. In order to form the duplex, the complementary strand must open this structure, and this is a competitive process in that the intramolecular hybridization that closes the stem must be displaced by hybridization of the complementary strand to the loop. As a result, hybridization can be very selective, and in optimal cases, single-base mismatch selectivity is observed (*8*, *9*). That is, the perfect complement hybridizes to the hairpin, but a strand containing even a single mismatch does not.

Our hairpin-DNA transporter (Table 1) was 30 bases long and contained a thiol substituent at the 5′ end that allowed it to be covalently attached to the inside walls of the gold nanotubes (*13*). The first six bases at each end of this molecule are complimentary to each other and form the stem of the hairpin, and the middle 18 bases form the loop (Table 1). The permeating DNA molecules were 18 bases long and were either perfectly complementary to the bases in the loop or contained one or more mismatches with the loop (Table 1). A second thiol-terminated DNA transporter was also investigated (Table 1). This DNA transporter was also 30 bases long, and the 18 bases in the middle of the strand were identical to the 18 bases in the loop of the hairpin-DNA transporter. However, this second DNA transporter did not have the complementary stem-forming bases on either end and thus could not form a hairpin. We used this linear-DNA transporter to test the hypothesis that the hairpin-DNA provides better transport selectivity because of its enhanced ability to discriminate the perfect-complement permeating DNA from the permeating DNAs that contained mismatches.

The transport experiments were done in a U-tube permeation cell (*10*) in which the gold nanotube membrane separated the feed half-cell containing one of the permeating DNA molecules (Table 1), dissolved in pH 7.2 phosphate buffer (ionic strength ~0.2 M), from the permeate half-cell that initially contained only buffer. We determined the rate of transport (flux) of the permeating DNA molecule from the feed half-cell through the membrane into the permeate half-cell by periodically measur-

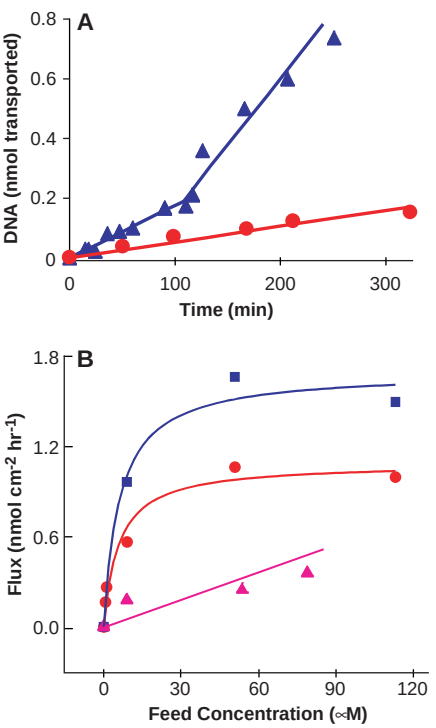


Fig. 1. (**A**) Transport plots for PC-DNA through gold nanotube membranes with (blue triangles) and without (red circles) the immobilized hairpin-DNA transporter. The feed solution concentration was 9 μM. (**B**) Flux versus feed concentration for PC-DNA. The data in red and blue were obtained for a gold nanotube membrane containing the hairpin-DNA transporter. At feed concentrations of 9 μM and above, the transport plot shows two linear regions. The data in blue (squares) were obtained from the high slope region at longer times. The data in red (circles) were obtained from the low slope region at shorter times. The data in pink (triangles) were obtained for an analogous nanotube membrane with no DNA transporter.

Department of Chemistry and Center for Research at the Bio/Nano Interface, University of Florida, Gainesville, FL 32611–7200, USA.

*To whom correspondence should be addressed. E-mail: crmartin@chem.ufl.edu

ing the ultraviolet absorbance of the permeate half-cell solution, at 260 nm, that arose from the permeating DNA molecule.

Transport plots (Figs. 1A and 2) show the number of nanomoles of the permeating DNA transported through the nanotube membrane versus permeation time. When the hairpin DNA was not attached, a straight-line transport plot was obtained for the perfect-complement DNA (PC-DNA) (Fig. 1A), and the slope of this line provides the flux of PC-DNA across the membrane (Table 2). The analogous transport plot for the membrane containing the hairpin-DNA transporter is not linear, but instead can be approximated by two straight-line segments: a lower slope segment at short times followed by a higher slope segment at times longer than a critical transition time, τ. This transition is very reproducible; for example, for a feed concentration of 9 μM, τ was 110 $\pm$ 15 min (average of three membranes).

Figure 1A shows that the flux of the permeating PC-DNA in the membrane containing the hairpin-DNA transporter was at all times higher than the flux for an otherwise identical membrane without the transporter (Table 2). Hence, the hairpin DNA acted as an MR agent to facilitate the transport (*4*, *5*, *14*, *15*) of the PC-DNA. Additional evidence for this conclusion was obtained from studies of the effect of concentration of the PC-DNA in the feed solution on the PC-DNA flux. If the hairpin DNA facilitated the transport of the PC-DNA, this plot should show a characteristic "Langmuirian" shape (*4*, *5*). Figure 1B shows that this is indeed the case, for transport data both before and after τ. The analogous plot for the identical membrane without the hairpin-DNA transporter is linear (Fig. 1B), showing that transport is not facilitated but rather described simply by Fick's first law of diffusion. The transition to the higher slope segment was not observed, during permeation experiments with a total duration of 300 min, for feed concentrations below 9 μM (Fig. 1B).

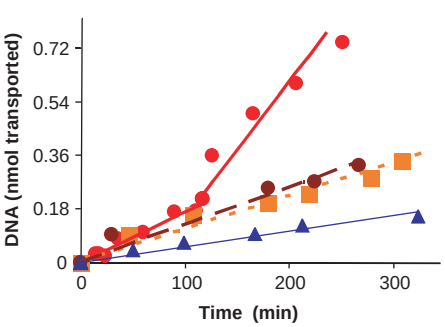


Fig. 2. Transport plots for a gold nanotube membrane containing the hairpin-DNA transporter. The permeating DNA was either PC-DNA (red circles), single-mismatch (end) (brown circles), seven-mismatch (blue triangles), or single-mismatch (middle) (orange squares). The feed solution concentration was 9 μM.

Analogous permeation data were obtained for the various mismatch-containing permeating DNA molecules (Table 1). The transport plots for these mismatch DNAs show only one straight-line segment (Fig. 2), and their fluxes were always lower than the flux for the PC-DNA obtained from the higher slope region of the PC-DNA transport plot (Table 2). In particular, the membrane containing the hairpin-DNA transporter showed higher flux for PC-DNA than for the two permeating DNAs that contained only a single-base mismatch.

To illustrate this point more clearly, we defined a selectivity coefficient $\alpha_{HP,PC/1MM}$, which is the flux for the PC-DNA divided by the flux for a single-base mismatch DNA in the membrane with the hairpin (HP)–DNA transporter. A selectivity coefficient of $\alpha_{HP,PC/1MM} = 3$ can be derived from the data in Table 1. The analogous selectivity coefficient for the PC-DNA versus the DNA with seven mismatches is $\alpha_{HP,PC/7MM} = 7$. These selectivity coefficients show that nanotube membranes containing the hairpin-DNA transporter selectively transport PC-DNA and that single-base mismatch transport selectivity can be obtained.

The importance of the hairpin structure to membrane selectivity is illustrated by analogous transport data for membranes containing the linear-DNA transporter (Table 1). With this transporter, all of the transport plots show only a single straight-line segment, and the fluxes for the single-mismatch DNAs were identical to the flux for the PC-DNA (Table 2); i.e., the single-base mismatch selectivity coefficient for this linear (LN) DNA transporter is $\alpha_{LN,PC/1MM} = 1$. The linear-DNA transporter does, however, show some transport selectivity for the PC-DNA versus the seven-mismatch DNA, $\alpha_{LN,PC/7MM} = 5$.

We also investigated the mechanism of transport in these membranes. In such MR-based, facilitated-transport membranes, the permeating species is transported by sequential binding and unbinding events with the MR agent (*4*, *5*, *14*). For these DNA-based membranes, the binding and unbinding events are sequential hybridization and dehybridization reactions between the permeating DNA molecule and the DNA transporter attached to the nanotubes. To show that hybridization occurred in the membrane with the hairpin-DNA transporter, the membrane was exposed to PC-DNA and then to a restriction enzyme (Sfc I, New England Biolabs) (*13*). If hybridization between the PC-DNA and the hairpin transporter occurs, this enzyme would cut the resulting double-stranded DNA such that the last five bases of the binding loop, and all of the stem-forming region, at the 3′ end of the hairpin would be removed. This reaction would substantially damage the binding site, and on the basis of our prior work (*4*), we predicted that if this membrane were subsequently used in a permeation experiment, a lower PC-DNA flux would be obtained (*16*).

After exposure to the restriction enzyme, the membrane was extensively rinsed to remove the enzyme and DNA fragments and was then

Table 1. DNA molecules used. For transporter DNAs, the 18 bases that bind to the permeating DNAs are in bold. For permeating DNAs, the mismatched bases are underlined. FAM is a fluorescein derivative (Applied Biosystems), and Cy5 is a cyanine dye (Amersham Biosciences).

Type	Sequence
Transporter DNAs	
Hairpin	5′HS-$(CH_2)_6$-CGCGAG**AAGTTACATGACCTGTAG**CTCGCG3′
Linear	5′HS-$(CH_2)_6$-CGCGAG**AAGTTACATGACCTGTAG**ACGATC3′
Permeating DNAs	
Perfect complement (PC-DNA)	3′TTCAATGTACTGGACATC5′
Single-base mismatch (3′ end)	3′C̲TCAATGTACTGGACATC5′
Single-base mismatch (middle)	3′TTCAATGTAG̲TGGACATC5′
Seven-mismatch	3′AAGTTACATGACCTGTAG5′
FAM-labeled perfect complement	3′TTCAATGTACTGGACATC-$(CH_2)_6$-FAM 5′
Cy5-labeled single-base mismatch	3′C̲TCAATGTACTGGACATC-$(CH_2)_6$ Cy5 5′

Table 2. Fluxes for a feed concentration of 9 μM.

Transporter DNA	Permeating DNA	Flux (nmol cm^{-2} h^{-1})
Hairpin	Perfect complement	0.57, 1.14*
Linear	Perfect complement	0.94
None	Perfect complement	0.20
Hairpin	Single mismatch (middle)	0.37
Hairpin	Single mismatch (end)	0.44
Linear	Single mismatch (middle)	0.94
Hairpin	Seven-mismatch	0.17
Linear	Seven-mismatch	0.20

*Two fluxes were obtained because the transport plot showed two slopes (Fig. 1A).

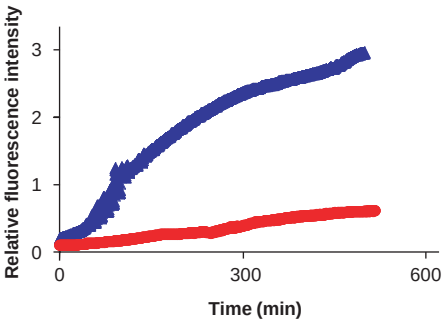


Fig. 3. Release of fluorescently labeled PC-DNA from a membrane containing the hairpin-DNA transporter. The fluorescently labeled PC-DNA was released into a buffer solution containing no unlabeled PC-DNA (lower curve) or into a buffer containing 9 μM unlabeled PC-DNA (upper curve).

used for a transport experiment with PC-DNA as the permeating species. Unlike the data in Fig. 1A, the transport plot for this damaged-transporter membrane showed only one straight-line segment (*13*), corresponding to a flux of 0.2 nmol cm^{-2} h^{-1}. This value is well below those we observed from membranes with an undamaged DNA-hairpin transporter (Table 2). The damaged DNA-transporter was then removed from the nanotubes, and fresh DNA-hairpin transporter was applied. A subsequent transport experiment with PC-DNA showed a transport plot identical to that obtained before exposure to the restriction enzyme (*13*). These data suggest that hybridization is, indeed, involved in the transport mechanism for the DNA-hairpin–containing membranes.

To show that dehybridization occurs on a reasonable time scale in these membranes, we exposed a hairpin-DNA membrane to a fluorescently labeled version of the PC-DNA (Table 1). The membrane was then rinsed with buffer solution and immersed into a solution of either pure buffer or buffer containing unlabeled PC-DNA. If the dehybridization reaction is facile, the fluorescently labeled PC-DNA should be released into the solution. We found that dehybridization did occur, but it was strongly accelerated when unlabeled PC-DNA was present in the solution (Fig. 3). Hence, dehybridization is much faster when it occurs through a cooperative process whereby one PC-DNA molecule displaces another from an extant duplex (*17*).

We also investigated transport selectivity for a feed solution containing fluorescently labeled versions (Table 1) of both the PC-DNA and the single-mismatch DNA. The fluorescent labels allowed for quantification of both of these permeating DNAs simultaneously in the permeate solution. In analogy to the single-molecule permeation experiment, the flux of the PC-DNA was five times higher than the flux of the single-mismatch DNA (*13*). To assess the practical utility of these membranes, transport studies with more realistic samples (such as cell lysates) will be needed.

Finally, we have not observed a spontaneous transition from a low-flux to a high-flux state (Fig. 1A) with our previous MR-based membranes (*4*, *5*). The fact that whether this transition is observed depends on the feed concentration suggests that the transition is a transport-related phenomenon. It is possible that this transition relates to the concept of cooperative (high-flux) versus noncooperative (low-flux) dehybridization (Fig. 3), but further studies, both experimental and modeling, will be required before a definitive mechanism for this transition can be proposed.

References and Notes

1. D. A. Doyle *et al.*, *Science* **280**, 69 (1998).
2. J. Abramson *et al.*, *Science* **301**, 610 (2003).
3. B. Hille, *Ion Channels of Excitable Membranes* (Sinauer, Sunderland, MA, ed. 3, 2001), pp. 441–470.
4. S. B. Lee *et al.*, *Science* **296**, 2198 (2002).
5. B. B. Lakshmi, C. R. Martin, *Nature* **388**, 758 (1997).
6. An interesting example of attaching a single-stranded DNA molecule to a protein channel to make a new type of DNA sensor has been reported (*18*).
7. H. Fried, U. Kutay, *Cell. Mol. Life Sci.* **60**, 1659 (2003).
8. G. Bonnet, S. Tyagi, A. Libchaber, F. R. Kramer, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6171 (1999).
9. B. Dubertret, M. Calame, A. J. Libchaber, *Nature Biotechnol.* **19**, 365 (2001).
10. C. R. Martin, M. Nishizawa, K. Jirage, M. Kang, *J. Phys. Chem. B* **105**, 1925 (2001).
11. K. B. Jirage, J. C. Hulteen, C. R. Martin, *Science* **278**, 655 (1997).
12. C. R. Martin, *Science* **266**, 1961 (1994).
13. Materials and methods are available as supporting material on *Science* Online.
14. M. Mulder, *Basic Principles of Membrane Technology* (Kluwer, Dordrecht, Netherlands, 1996), pp. 342–351.
15. Y. Osada, T. Nakagawa, in *Membrane Science and Technology*, Y. Osada, T. Nakagawa, Eds. (Marcel Dekker, New York, 1992), pp. 377–391.
16. A fluorescence-based method was used to provide direct evidence for clipping of the double-stranded DNA by the restriction enzyme (*13*).
17. M. C. Hall, H. Wang, D. A. Erie, T. A. Kunkel, *J. Mol. Biol.* **312**, 637 (2001).
18. S. Howorka, S. Cheley, H. Bayley, *Nature Biotech.* **19**, 636 (2001).
19. Supported by the National Science Foundation and by the Defense Advanced Research Projects Agency.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/984/DC1
Materials and Methods
Figs. S1 to S3
References and Notes

6 May 2004; accepted 9 July 2004

Sample Dimensions Influence Strength and Crystal Plasticity

Michael D. Uchic,[1]* Dennis M. Dimiduk,[1] Jeffrey N. Florando,[2] William D. Nix[3]

When a crystal deforms plastically, phenomena such as dislocation storage, multiplication, motion, pinning, and nucleation occur over the submicron-to-nanometer scale. Here we report measurements of plastic yielding for single crystals of micrometer-sized dimensions for three different types of metals. We find that within the tests, the overall sample dimensions artificially limit the length scales available for plastic processes. The results show dramatic size effects at surprisingly large sample dimensions. These results emphasize that at the micrometer scale, one must define both the external geometry and internal structure to characterize the strength of a material.

A size-scale effect can be defined as a change in material properties—mechanical, electrical, optical, or magnetic—that is due to a change in either the dimensions of an internal feature or structure or in the overall physical dimensions of a sample. For metals, size-scale effects related to changes in internal length scales are readily observed and are often exploited for industrial use. For example, it is well known that the yield strengths of metallic alloys can be improved through refinement of the grain size (*1–3*), where the yield strength is proportional to the inverse square root of the average grain diameter, and this relation is generally valid for grains that range in size from millimeters to tens of nanometers. By comparison, changes in the mechanical response of materials due solely to the physical geometry of a sample have been largely overlooked. Large increases in yield strength (approaching the theoretical limit) were observed over 40 years ago in tension testing of single-crystal metallic whiskers having micrometer-scale diameters (*4–6*). However, whisker testing is restricted to materials that can be grown in that form. Conversely, no changes in strength and only mild decreases in work hardening were observed during the deformation of

[1]Air Force Research Laboratory, Materials & Manufacturing Directorate, Wright-Patterson Air Force Base, OH 45433–7817, USA. [2]Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. [3]Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305–2205, USA.

*To whom correspondence should be addressed. E-mail: michael.uchic@wpafb.af.mil

simple metals at submillimeter sample diameters (*7–10*), but those studies only started to explore the gap between millimeter and whisker dimensions.

There remains a fundamental challenge to systematically investigate external length-scale effects in the submillimeter-to-nanometer size regime. Such small dimensions are pervasive in modern devices and also encompass the size range in which dislocation-based plasticity mechanisms occur. External length-scale effects may be observed at multiple stages over this wide range of sizes, because the mechanisms associated with dislocation storage, multiplication, motion, pinning, and nucleation are generally active over different length scales. Without such an understanding, it is impossible to know the appropriate material properties to use in the design of small devices. At present, one can question whether features having micrometer-sized dimensions should be designed using the extraordinary strengths of defect-free "whiskers" or using behavior more akin to that of bulk metal crystals.

Recently, size-scale effects in materials mechanics received renewed attention under conditions where deformation gradients are imposed at the micrometer scale (*11–14*). These studies explore the evolution of geometrically necessary dislocations (GNDs) (*15*, *16*) that are required to accommodate the plastic strain gradients that may be induced by the test condition or by the internal structure of the material. For example, the permanent change in the profile of a surface during indentation testing, which is due to the deformation gradient imposed by the indentation tip, may be wholly accommodated by the generation and motion of GNDs. These studies find that gradient-induced increases in defect evolution result in concomitant local changes in the strength and hardening rates of materials. However, the studies do not consider other changes in the fundamental deformation mechanisms associated with limiting the physical dimensions of the deforming volume; that is, one might speculate that the deformation micromechanisms themselves are affected by the size of the deforming volume. We suggest that a more complete understanding of size effects for a given material can only be realized after testing for geometric effects and under test conditions that minimize imposed deformation gradients, thus limiting the GND density.

We have developed a test methodology (*17*) that allows the exploration of size-scale effects in virtually any bulk inorganic material, using a focused ion beam (FIB) microscope for sample preparation, together with mechanical testing that is a simple extension of nanoindentation technology. The FIB is used to machine cylindrical compression samples into the surface of a bulk crystal, leaving the samples attached to the bulk at one end. Samples were prepared in the size range from 0.5 to 40 μm in diameter and with an aspect ratio ranging from 2:1 to 4:1. Once prepared, the samples were tested using a conventional nanoindentation device outfitted with a flat-punch indentation tip. Nanoindentation systems are normally used for depth-sensing indentation experiments using a sharp tip, but here the same platform is used to perform conventional uniaxial compression tests at prescribed displacement rates ranging from 1 to 5 nm/s. This technique can be used to study external size effects in single crystals in the absence of grain boundaries, which are strong internal barriers to dislocation glide. Although there have been notable advances in mechanical test methods that operate on micrometer-sized samples (*18–20*), these test techniques use samples that have been fabricated with wafer processing methods that are specific to the microelectronic industry. The microstructures of those samples are typically polycrystalline, having a submicrometer grain size, which can complicate the interpretation of observed external size effects (*21*).

The mechanical behavior of bulk single crystals of pure Ni is well known, so this material was used as a model system for the test method. A single-slip orientation was selected to simplify the defect evolution and hardening conditions. The stress-strain curves for Ni microcompression samples having diameters in the 20- to 40-μm range are similar to those for bulk samples (Fig. 1A), because the yield strength and overall work-hardening rates are within 30% of the measured properties of millimeter-sized specimens. After testing, fine discrete slip bands can be observed along the gauge length of the samples, which are also found in the bulk specimen tests (Fig. 1B).

For samples 5 μm in diameter, there are distinct changes in the stress-strain curves that are indicative of physical size limitations. These samples display large strain bursts: very rapid flow to values up to 19% strain upon yielding, in contrast to bulk samples that show a smooth transition from elastic to plastic flow and a steady rate of work hardening. The yield stress for each of the four 5-μm-diameter samples is higher than that for microsamples that are equal to or larger than 20 μm in diameter and varies over a range of 70 MPa. Differences may also be observed in the appearance of the microsamples after testing. There are fewer slip bands, but those that exist appear to be much more active, as indicated by large single-slip plane displacements (Fig. 1C). Strain bursts are also observed for samples 10 μm in diameter, although the

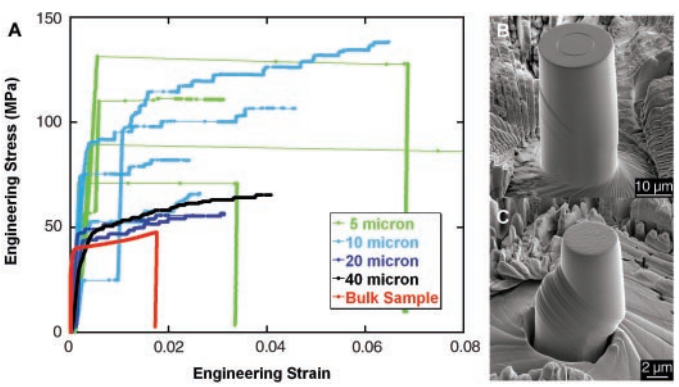


Fig. 1. Mechanical behavior at room temperature for pure Ni microsamples having a <134> orientation. (**A**) Stress-strain curves for microsamples ranging in size from 40 to 5 μm in diameter, as well as the stress-strain curve for a bulk single crystal having approximate dimensions 2.6 × 2.6 × 7.4 mm. (**B**) A scanning electron micrograph (SEM) image of a 20-μm-diameter microsample tested to ~4% strain. The circle milled into the top surface of the microsample is a fiducial mark used during sample machining. (**C**) A SEM image of 5-μm-diameter microsample after testing, where the sample achieved ~19% strain during a rapid burst of deformation that occurred in less than 0.2 s.

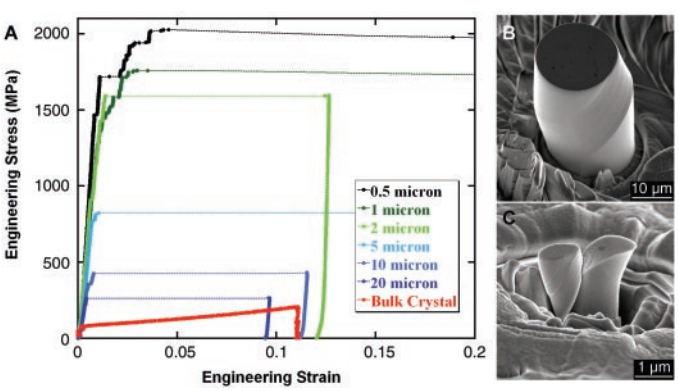


Fig. 2. Mechanical behavior at room temperature for Ni_3Al-Ta microsamples having a <123> orientation. (**A**) Representative stress-strain curves for microsamples ranging in size from 20 to 0.5 μm in diameter, as compared to the behavior of a bulk single crystal having approximate dimensions 2.5 × 2.5 × 7.5 mm. (**B**) A SEM image of 20-μm-diameter microsample after testing, where the sample achieved ~10% strain during the rapid burst of deformation. (**C**) A SEM image of 1-μm-diameter microsample after testing, where the top of the sample has completely sheared off during the rapid strain burst. This behavior is observed for both the 1- and 0.5-μm-diameter samples.

extent of these is typically less than 1% strain. For samples 10 μm in diameter and larger, most of the plastic deformation consists of short periods of stable flow with low work-hardening rates, separated by increments of nearly elastic loading. There is a gradual progression between bulk and size-limited behavior as the sample size decreases from 40 to 5 μm in diameter. These attributes are distinct from the common behavior of both bulk materials and whiskers. Whiskers of pure metals typically display much higher yield stresses than bulk materials. In one study, the strength of Cu whiskers 16 μm in diameter and smaller exhibited yield stresses in the range from 0.3 to 6 GPa (*6*), whereas the yield stress for bulk Cu is on the order of 10 to 50 MPa (depending on purity levels and heat treatment conditions). In addition, after yielding, whiskers do not maintain this high flow stress; rather, the flow stress drops to the level observed in bulk Cu or the whisker simply fractures. The reasons for this are understood to be related to the fact that, unlike most common metals, the whiskers start out defect-free before loading.

One interpretation of these results is that decreasing sample diameter affects the mechanisms for defect multiplication and storage that are associated with plastic flow, before the dislocation-source–limited regime attributed to whiskers is achieved. The increases in flow stress and extremely low hardening rates fall outside the regimes known for bulk tests but do not enter the regime of high stresses known for metal whiskers. The increase in the spread and the rise of the values of the yield stress for smaller samples suggest aspects of self-organization and criticality events at the elastic-plastic transition. That is, the transition appears to be stochastic, showing a progression toward a single catastrophic event as the ability to multiply dislocations or the number of dislocation sources is truncated. This occurs either through increasing levels of deformation or through shrinking the total volume of the sample.

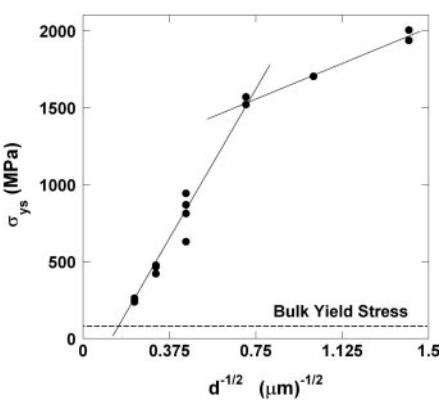


Fig. 3. Dependence of the yield strength on the inverse of the square root of the sample diameter for Ni_3Al-Ta. The linear fit to the data predicts a transition from bulk to size-limited behavior at ~42 μm. σ_{ys}, the stress for breakaway flow.

The same method was used to examine an intermetallic alloy, Ni_3Al-Ta, which is widely known to exhibit fundamentally different flow mechanisms. One physical manifestation of this behavior is an anomalous increase in strength with increasing temperature. There is considerable evidence that at temperatures in the anomalous flow regime, the mobility of screw-character dislocations is greatly influenced by the lateral motion of large jogs and kinks along the length of the dislocations (*22*–*24*), and it is likely that dislocation kinetics are strongly influenced by the characteristic active line length of dislocations known to be on the order of a few micrometers (*23*, *25*). The characteristic scales for multiplication are unknown. In the present study, the sample sizes are equivalent to the length scales for the physical processes governing flow.

We observed a dramatic size effect on strength for a Ni_3Al-1% Ta alloy deforming under nominally single-slip conditions (Fig. 2A). The flow stress increased from 250 MPa for a 20-μm-diameter sample to 2 GPa for a 0.5-μm-diameter sample. These flow stresses are much higher than those found for bulk crystals, which themselves exhibit a flow stress of only 81 MPa. Although these stresses exceed those for the bulk material, the influence of sample size occurs at dimensions that are large by comparison to whisker-type tests. After testing, slip traces are very fine and are homogeneously distributed along the gage section (Fig. 2B), except for the 0.5- and 1-μm-diameter samples, because they have completely sheared apart during large strain bursts. Closer inspection of the loading curves for all of the tests before the large strain bursts show small events of plastic activity that occur sporadically during the loading of the sample, separated by nearly elastic loading, again akin to self-organized processes. These aspects of work-hardening behavior are similar to what we have observed in the smaller Ni samples but have not been reported for bulk samples.

Examination of the flow stress in Ni_3Al-Ta as a function of sample diameter (Fig. 3) shows two regimes of size-dependent strengthening that scale with the inverse of the square root of the sample diameter—coincidently similar to grain-size hardening. However, although such strength scaling in metals usually arises from the presence of internal kinematic barriers to flow, these samples have no known internal barriers. One may speculate that this remarkable behavior is associated with changes in the self-exhaustion or annihilation of dislocations, specifically those of screw character. That said, it is surprising that significant length-scale effects are observed for such large sample sizes; note that the transition to bulk behavior is predicted from the scaling relation in Fig. 3 to occur for samples greater than 42 μm in diameter.

Finally, we examined a Ni superalloy single crystal that consisted of a Ni solid-solution matrix having a high volume fraction of Ni_3Al-based precipitates that are ~250 nm in diameter and are uniformly distributed. Both solid-solution alloying and the precipitates provide additional strengthening mechanisms and help to determine internal deformation length scales. A 10-μm-diameter microcompression sample, which had about 30 precipitates spanning the width of the sample, displayed a mechanical response that matched the behavior of a bulk tension test (Fig. 4). The agreement is not surprising, because the strong internal hardening mechanisms that control plastic deformation operate at the dimensional scale of the precipitates and are still effective at this sample size, thus preempting influences from limited sample dimensions.

We have demonstrated a method to characterize aspects of length-scale effects on deformation and strength by shrinking the traditional uniaxial compression test to the micrometer scale. From these tests it is clear that when the external dimensions of the

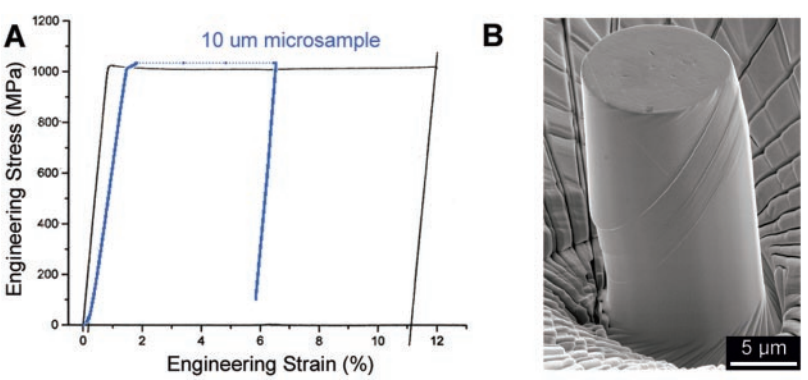


Fig. 4. Mechanical behavior at room temperature of a Ni superalloy microsample having a near-<001> orientation. (**A**) A stress-strain curve for a 10-μm-diameter microsample tested in compression as compared to the behavior of a bulk single tested in tension. The microsample was machined from an undeformed region of the grip region of the bulk sample after testing. (**B**) A SEM image of the microsample after testing.

sample become smaller than a few tens of micrometers, the basic processes of plastic deformation are affected; thus, it may not be possible to define the strength of a given material in the absence of physical conditions that are completely specified. The results show that such influences occur at much larger dimensions than are classically understood for metal whisker-like behavior (*6*). Emerging strain-gradient–based continuum theories of deformation (that is, models that incorporate a physical length scale into the constitutive relations for the mechanical response of materials) must carefully account for these fundamental changes of deformation mechanisms that extend beyond the gradient-induced storage of defects.

References and Notes

1. E. O. Hall, *Proc. Phys. Soc. London B* **64**, 747 (1951).
2. N. J. Petch, *J. Iron Steel Inst.* **174**, 25 (1953).
3. S. Yip, *Nature* **391**, 532 (1998).
4. S. S. Brenner, *J. Appl. Phys.* **27**, 1484 (1956).
5. S. S. Brenner, *J. Appl. Phys.* **28**, 1023 (1957).
6. S. S. Brenner, in *Growth and Perfection of Crystals*, R. H. Doremus, B. W. Roberts, D. Turnbull, Eds. (Wiley, New York, 1959), pp. 157–190.
7. H. Suzuki, S. Ikeda, S. Takeuchi, *J. Phys. Soc. Jpn.* **11**, 382 (1956).
8. J. T. Fourie, *Philos. Mag.* **17**, 735 (1968).
9. S. J. Bazinski, Z. S. Bazinski, in *Dislocations in Solids*, F. R. N. Nabarro, Ed. (North Holland, Amsterdam, 1979), vol. 4, pp. 261–362.
10. G. Sevillano, in *Materials Science and Technology, Vol. 6 Plastic Deformation and Fracture of Materials*, H. Mughrabi, Ed. (VCH, Weinheim, Germany, 1993), vol. 6, pp. 19–88.
11. N. A. Fleck, G. M. Muller, M. F. Ashby, J. W. Hutchinson, *Acta Metall. Mater.* **42**, 475 (1994).
12. Q. Ma, D. R. Clarke, *J. Mater. Res.* **10**, 853 (1995).
13. W. D. Nix, H. Gao, *J. Mech. Phys. Solids* **46**, 411 (1998).
14. J. S. Stölken, A. G. Evans, *Acta Mater.* **46**, 5109 (1998).
15. J. F. Nye, *Acta Metall.* **1**, 153 (1953).
16. M. F. Ashby, *Philos. Mag.* **21**, 399 (1970).
17. M. D. Uchic, D. M. Dimiduk, J. N. Florando, W. D. Nix, in *Materials Research Society Symposium Proceedings*, E. P. George *et al.*, Eds. (Materials Research Society, Pittsburgh, PA, 2003), vol. 753, pp. BB1.4.1–BB1.4.6.
18. W. N. Sharpe, K. M. Jackson, K. J. Hemker, Z. Xie, *J. MEMS Syst.* **10**, 317 (2001).
19. M. A. Haque, M. T. A. Saif, *Sensors Actuators A* **97-98**, 239 (2002).
20. H. D. Espinosa, B. C. Prorok, M. Fischer, *J. Mech. Phys. Solids* **51**, 47 (2003).
21. H. D. Espinosa, B. C. Prorok, B. Peng, *J. Mater. Res.* **52**, 667 (2004).
22. M. Mills, N. Baluc, H. P. Karnthaler, in *Materials Research Society Symposium Proceedings*, C. T. Liu *et al.*, Eds. (Materials Research Society, Pittsburgh, PA, 1989), vol. 133, pp. 203–208.
23. P. Veyssière, G. Saada, in *Dislocations in Solids*, F. R. N. Nabarro, M. S. Duesbery, Eds. (North Holland, Amsterdam, 1996), vol. 10, pp. 253–440.
24. P. B. Hirsch, *Philos. Mag. A* **65**, 569 (1992).
25. X. Shi, G. Saada, P. Veyssière, *Philos. Mag. Lett.* **71**, 1 (1995).

26. Supported by the Air Force Office of Scientific Research and the Accelerated Insertion of Materials program of the Defense Advanced Research Projects Agency (M.D.U. and D.M.D.) under the direction of C. Hartley and L. Christodoulou, respectively, and by the U.S. Department of Energy and NSF (J.N.F. and W.D.N.). The Ni superalloy single crystal and corresponding bulk mechanical test data were provided by T. Pollock of the University of Michigan. We gratefully acknowledge useful discussions with T. A. Parthasarathy, K. Hemker, and R. LeSar. We also acknowledge H. Fraser, whose efforts enabled many aspects of the instruments used in this work.

9 April 2004; accepted 21 July 2004

Discovery of Mass Anomalies on Ganymede

John D. Anderson,[1]* Gerald Schubert,[2,3] Robert A. Jacobson,[1] Eunice L. Lau,[1] William B. Moore,[2] Jennifer L. Palguta[2]

We present the discovery of mass anomalies on Ganymede, Jupiter's third and largest Galilean satellite. This discovery is surprising for such a large icy satellite. We used the radio Doppler data generated with the Galileo spacecraft during its second encounter with Ganymede on 6 September 1996 to model the mass anomalies. Two surface mass anomalies, one a positive mass at high latitude and the other a negative mass at low latitude, can explain the data. There are no obvious geological features that can be identified with the anomalies.

Jupiter's four Galilean satellites can be approximated by fluid bodies that are distorted by rotational flattening and by a static tide raised by Jupiter. All four satellites are in synchronous rotation with their orbital periods, and all four are in nearly circular orbits in Jupiter's equatorial plane (*1*). Previously we reported on the interior structure of the four Galilean satellites as inferred from their mean densities and second-degree (quadrupole) gravity moments (*2*). The inner three satellites, Io, Europa, and Ganymede, have differentiated into an inner metallic core and an outer rocky mantle. In addition, Europa and Ganymede have deep icy shells on top of their rocky mantles. The outermost satellite, Callisto, is an exception. It has no metallic core, and rock (plus metal) and ice are mixed throughout most if not all of its deep interior.

These interior models are consistent with the satellites' external gravitational fields, as inferred from radio Doppler data from close spacecraft flybys, with one exception. It is impossible to obtain a satisfactory fit to the Doppler data from the second Ganymede flyby (G2) without including all gravity moments to the fourth degree and order in the fitting model. The required truncated spherical harmonic expansion for Ganymede's gravitational potential function V takes the form (*3*)

$$V(r,\phi,\lambda) = \frac{GM}{r}\left[1+\sum_{n=2}^{4}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{n}(C_{nm}\cos m\lambda + S_{nm}\sin m\lambda)\,P_{nm}(\sin\phi)\right] \quad (1)$$

The spherical coordinates (r, ϕ, λ) are referred to the center of mass, with r the radial distance, ϕ the latitude, and λ the longitude on the equator. P_{nm} is the associated Legendre polynomial of degree n and order m, and C_{nm} and S_{nm} are the corresponding harmonic coefficients, determined from the Doppler data by least squares analysis.

By taking the satellite's center of mass at the origin, its first degree coefficients are zero by definition, although all harmonics of degree greater than one can be nonzero (*3*). The gravity parameters needed to fit the G2 Doppler data to the noise level are GM, where M is the mass of the satellite and G is the gravitational constant; five second-degree coefficients; seven third-degree coefficients; and nine fourth-degree coefficients, for a total of 22 gravity parameters. With only two flybys and no global coverage of Ganymede's gravitational potential, the truncated harmonic expansion is not unique. Consequently, the 22 gravity parameters have little physical meaning.

The reference radius R for the potential function is set equal to the best determination of Ganymede's physical radius from spacecraft images, 2631.2 ± 1.7 km (*4*). With this radius, Ganymede is the largest satellite in the solar system, larger than Saturn's satellite Titan and even larger than the planet Mercury. Its GM value, as determined by four flybys (G1, G2, G7, G29), is 9887.83 ± 0.03 km^3 s^{-2} (*4*), which yields a mean density of 1941.6 ± 3.8 kg m^{-3}, consistent with a differentiated metal-rock interior and an icy shell about 800 km deep (*2*). Ganymede's total mass is (1.48150 ± 0.00022) × 10^{23} kg, where the error is dominated by the uncertainty in G (*5*), not the uncertainty in GM.

The higher degree coefficients required to fit the Ganymede G2 data must be a reflection of some other, and more localized, distortion of the gravitational field. In order to describe this more localized field, we first obtained a best fit to Ganymede's global field with just the parameters GM and the five second-degree gravity coefficients. Using this field, we calculated Doppler residuals about the best fit. Before any

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109–8099, USA. [2]Department of Earth and Space Sciences, University of California, Los Angeles, CA 90095–1567, USA. [3]Institute of Geophysics and Planetary Physics, Los Angeles, CA 90095–1567, USA.

*To whom correspondence should be addressed; E-mail: john.d.anderson@jpl.nasa.gov

fitting model was applied, the structure of the Doppler data was dominated by the *GM* term (Fig. 1). After removal of the best-fit model for *GM* and the second-degree field, the residuals were dominated by the localized gravity anomaly or anomalies (Fig. 2). These residuals were numerically differentiated by the cubic-spline technique developed for lunar mascons (*6*), thereby yielding acceleration data along the line of sight (Fig. 3).

Mass points were moved around on the surface until the acceleration data were fit with an inverse square Newtonian acceleration on the spacecraft in free fall (Gm/d^2), with m the mass of a mass point fixed in the body of Ganymede and d its distance from the spacecraft as a function of time. This Newtonian acceleration was then projected on the line of sight in order to produce a model for the observed acceleration. Nonlinear least squares analysis was used to find the best fit for the masses and for the locations of the mass points.

A reasonably good fit to the acceleration data can be achieved with just two masses, although a better fit is achieved with three masses (Table 1 and Fig. 3). This suggests that there are at least two distinct gravity anomalies on Ganymede. The first can be represented by a positive mass of about 2.6×10^{-6} the mass of Ganymede on the surface at high latitude near the closest approach point, and the second by a negative mass of about 5.1×10^{-6} times the mass of Ganymede at low latitude. The first mass is needed to fit the positive peak in the acceleration data at closest approach. The second mass fits the peak after closest approach and fills in the large depression in the acceleration data, thereby providing a better overall fit. A third, smaller positive mass of about 8.2×10^{-7} the mass of Ganymede improves the overall fit and produces a better fit to the acceleration data just before closest approach (Table 1 and Fig. 3). Because these results and standard errors are obtained by formal nonlinear least squares analysis, the results are model dependent with three independent variables (mass, latitude, longitude) for each anomaly. The results do not necessarily imply that the physical anomalies are known to a similar accuracy.

The results of Table 1 indicate that the first two larger masses are stable against changes to the fitting procedure, but that their locations can change appreciably. This can be demonstrated by changing the starting conditions for the fit from mass values near the two solutions of Table 1 to mass values of zero. The least-squares procedure converges to the same solutions regardless of the starting values for the masses. This suggests that the solutions represent the best global fit to the data, at least for small masses placed on the surface. The values of the masses are stable to within a standard deviation, although the locations can change by tens of a standard deviation. There are no obvious geological structures at the locations of the mass anomalies on Ganymede's surface that could be identified as the sources of the anomalies.

There may be additional gravity anomalies on Ganymede, but they are undetectable with only the two close flybys available. There may also be gravity anomalies on other Galilean satellites, especially on Europa, which has a differentiated structure similar to that of Ganymede. The only Europa flyby suitable for anomaly detection is the one on the 12th orbital revolution

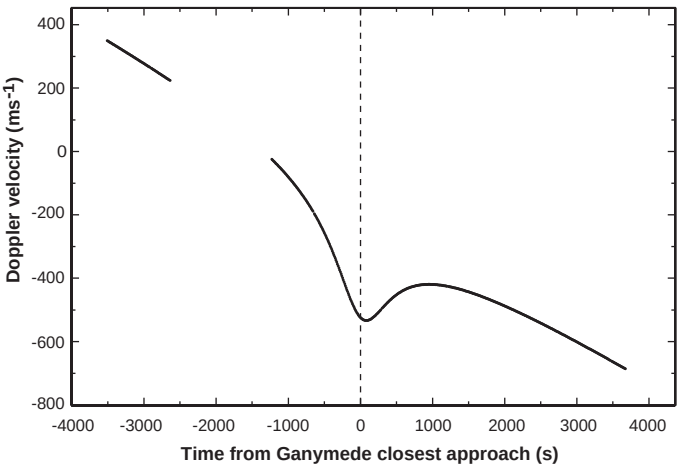


Fig. 1. Radio Doppler residuals before the application of any fitting model. The time tags for the raw Doppler data are in seconds from J2000 (JD 2451545.0 UTC) as measured by the station clock. The time tags for the plot are referenced to the G2 closest approach time of 6 September 1996, 19:38:34 UTC, ground receive time. The gap in the plot before closest approach is a result of a failure of the spacecraft receiver to phase lock to the uplink radio carrier wave. A reference frequency of 2.296268568 GHz has been subtracted from the raw data, and the result has been converted to Doppler velocity by the unit conversion factor Hz = 0.065278 m s^{-1}. The data are generated by sending a radio wave to the spacecraft, which returns it to the station by means of a radio transponder (two-way Doppler). Therefore the frequency reference for the Doppler shift is a hydrogen maser at the station, not the spacecraft's crystal oscillator. The sample interval for the data is 10 s.

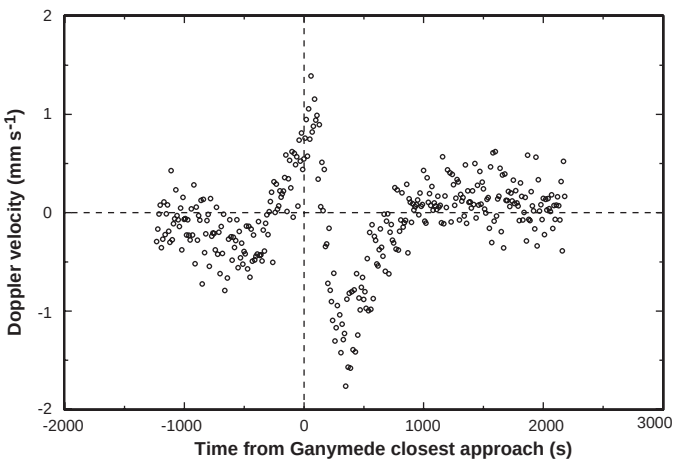


Fig. 2. Doppler residuals of Fig. 1, after the application of a fitting model that includes Ganymede's mass (*GM*) and its second degree and order gravity field. The remaining residuals are evidence of one or more gravity anomalies near the Galileo trajectory track.

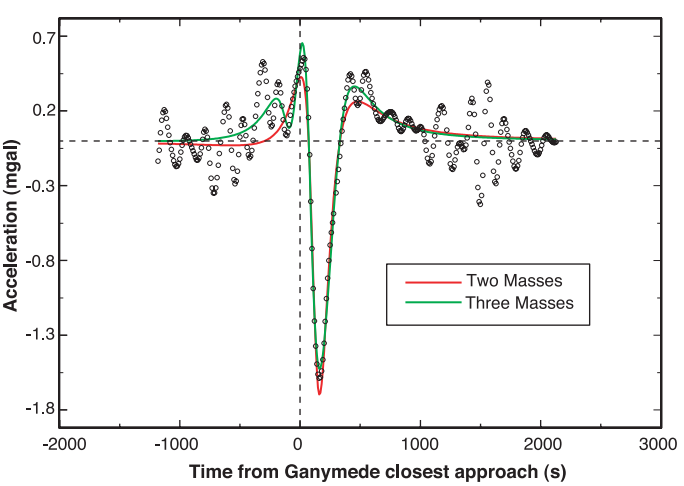


Fig. 3. Acceleration data in units of mgal (10^{-5} m s^{-2}) as derived from the Doppler residuals of Fig. 2. The best-fit acceleration model for two surface masses is shown in red. The best-fit model for three surface masses is in green.

Table 1. Least-squares fits to the acceleration data for two masses on Ganymede's surface and also for three masses on the surface. The three independent variables in the fitting model for each mass are *Gm*, and the geographic coordinates latitude and west longitude. For reference, the closest approach location is at latitude 79.3° and west longitude 123.7° at an altitude of 264 km. The measure of goodness of fit is given by the variance σ^2 for the acceleration residuals. A qualitative measure of the goodness of fit is given by Fig. 3.

	Six-parameter fit for two masses (σ^2 = 0.0244 mgal^2)		
	First mass	Second mass	Third mass
Gm (km^3 s^{-2})	0.0237 ± 0.0056	−0.0558 ± 0.0084	–
Latitude (°)	58.9 ± 1.5	24.2 ± 5.5	–
Longitude W (°)	65.2 ± 1.6	61.8 ± 5.4	–
	Nine-parameter fit for three masses (σ^2 = 0.0192 mgal^2)		
	First mass	Second mass	Third mass
Gm (km^3 s^{-2})	0.0256 ± 0.0038	−0.0500 ± 0.0058	0.0081 ± 0.0021
Latitude (°)	77.7 ± 1.0	39.9 ± 2.6	53.6 ± 2.3
Longitude W (°)	337.3 ± 5.1	355.6 ± 4.6	140.1 ± 4.8

(E12) at an altitude of 201 km. The E12 closest approach point is near the equator at a latitude of −8.7° and a west longitude of 225.7°. However, unlike G2 at an altitude of 264 km, Doppler data from E12, as well as three other more distant flybys (E4 at 692 km, E6 at 586 km, and E11 at 2043 km), can be fit to the noise level with second-degree harmonics. The two Callisto flybys that yield gravity information are more distant (C10 at 535 km and C21 at 1048 km). No anomalies are required to fit data from four Io flybys (I24 at 611 km, I25 at 300 km, I27 at 198 km, and I33 at 102 km). A satisfactory fit can be achieved with a second degree and order harmonic expansion for all the satellite flybys except G2, and for that one flyby even a third degree and order expansion leaves systematic Doppler residuals. The G2 flyby is unique.

The surface mass-point model provides a simple approach to fitting the data. Further analysis will be required to determine if other mass anomalies at different locations and depths below the surface might also yield acceptable fits to the Doppler residuals. Our fitting model of point masses does not allow specification of the horizontal dimensions over which the density heterogeneities extend, although these are likely to be hundreds of kilometers, comparable to the distances from the anomalies to the spacecraft. With additional study of the point-mass model and incorporation of more realistic anomaly shapes (disks, spheres) into the analysis, it may be possible to identify the physical sources of the anomalies. If the anomalies are at the surface, or near to it, then they could be supported for a lengthy period of geological time by the cold and stiff outer layers of Ganymede's ice shell.

References and Notes

1. A compilation of satellite data can be found in D. J. Tholen, V. G. Tejfel, A. N. Cox, *Allen's Astrophysical Quantities*, A. N. Cox, Ed. (Springer-Verlag, New York, ed. 4, 2000), pp. 302–310.
2. The interior composition, structure, and dynamics of the four Galilean satellites have been summarized, along with a bibliography, by G. Schubert, J. D. Anderson, T. Spohn, W. B. McKinnon, in *Jupiter*, F. Bagenal, T. E. Dowling, W. B. McKinnon, Eds. (Cambridge Univ. Press, New York, 2004), chap. 13.
3. W. M. Kaula, *Theory of Satellite Geodesy* (Blaisdell, Waltham, MA, 1966).
4. J. D. Anderson *et al*., Galileo Gravity Science Team, *Bull. Am. Astron. Soc.* **33**, 1101 (2001). This reference includes the best determination of Ganymede's radius currently available.
5. P. J. Mohr, B. N. Taylor, *Phys. Today* **55**, BG6 (2002). Because of recent determinations, the adopted value of *G* has fluctuated over the past few years. We use the current (2002) value recommended by the Committee on Data for Science and Technology (CODATA), $G = 6.6742 \times 10^{-11}$ m^3 kg^{-1} s^{-2}, with a relative standard uncertainty of 1.5×10^{-4}.
6. P. M. Muller, W. L. Sjogren, *Science* **161**, 680 (1968).

7. We acknowledge the work of Ö. Olsen for finding fits to the fourth-degree gravitational field. We thank W. L. Sjogren, A. S. Konopliv, and D.-N. Yuan for their assistance, especially for providing us with a recompiled version of their gravity-anomaly software *Gravity Tools*. We also thank D. Sandwell for helpful discussions about the nature of Ganymede's gravity anomalies, and S. Asmar, G. Giampieri, and D. Johnston for helpful discussions. This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA. G.S., W.B.M., and J.L.P. acknowledge support by grants from NASA through the Planetary Geology and Geophysics program.

12 April 2004; accepted 8 July 2004

Probing the Accumulation History of the Voluminous Toba Magma

Jorge A. Vazquez[1]*† and Mary R. Reid[1,2]

The age and compositional zonation in crystals from the Youngest Toba Tuff record the prelude to Earth's largest Quaternary eruption. We used allanite crystals to date and decipher this zoning and found that the crystals retain a record of at least 150,000 years of magma storage and evolution. The dominant subvolcanic magma was relatively homogeneous and thermally stagnant for ~110,000 years. In the 35,000 years before eruption, the diversity of melts increased substantially as the system grew in size before erupting 75,000 years ago.

Toba caldera, a continental arc volcano in Sumatra, Indonesia, produced Earth's largest Quaternary eruption, ejecting >3000 km^3 of magma 73,000 ± 4000 years ago (*1*). Atmospheric loading by aerosols and ash from the Toba eruption may have accelerated cooling of Earth's climate (*2*) and resulted in near-extinction of humans (*3*). How quickly this and other huge volumes of magma can amass is unclear, especially because large volumes of eruptible magma have not been detected beneath areas of active and/or long-lived magmatism (*4*, *5*). The rate of magma accumulation can dictate whether reservoirs of magma simply cool and solidify or persist at magmatic conditions (*6*, *7*), and may influence the probability of volcanic eruption and the characteristics of associated plutonic intrusions (*8*, *9*). A detailed record of magmatic evolution is that retained by the compositional zoning of major minerals (*10*, *11*), and this might reveal how magma chambers accumulate and change (*12*, *13*). However, current analytical techniques are not sufficiently sensitive to put the chemical zoning in major minerals into an absolute time frame. Hence, it is impossible to relate the zoning stratigraphy of one crystal to another or evaluate the age of magma associated with crystallization. Here we use a combination of in situ compositional and isotopic analyses on single crystals of a less abundant mineral, the epidote-group mineral allanite, to date and quantify compositional zoning within and between crystals in the Youngest Toba Tuff (YTT) and to establish how this voluminous magma evolved before eruption.

Allanite is a common accessory mineral in rhyodacitic and rhyolitic magmas and may have considerable compositional zoning in major and

[1]Department of Earth and Space Sciences, University of California, Los Angeles, CA 90095–1567, USA. [2]Department of Geology, Northern Arizona University, Flagstaff, AZ 86011, USA.

*To whom correspondence should be addressed. E-mail: jvazquez@ess.ucla.edu

†Present address: Department of Geological Sciences, California State University, Northridge, CA 91330–8266,USA.

minor elements. High Th concentrations (1 to 2 weight %) and a large degree of U-Th fractionation between allanite and melt make it ideal for in situ dating by ^{238}U-^{230}Th disequilibrium methods, with an age resolution of tens of thousands of years. Whereas individual zircons are also amenable to in situ dating (*14*), the composition of allanite is particularly sensitive to the differentiation of magma. The YTT is compositionally zoned from 68 to 77 weight % SiO_2, with the majority (>70%) of erupted magma being >73 weight % SiO_2 (*15*). Chesner (*15*) concluded that this diversity of compositions was largely produced by crystal fractionation. We analyzed allanites from a representative 75 weight % SiO_2 rhyolite with the UCLA high-resolution ion microprobe and an electron microprobe (*16*).

When the ^{238}U-^{230}Th isotope characteristics of the host rhyolite are used to estimate initial $^{230}Th/^{232}Th$ activity ratios, the cores of the YTT allanites are found to have crystallization ages ranging from 100 to 225 thousand years ago (ka), and most rims have ages identical to or within analytical error of the ~75-ka eruption age (*17*). Allanite compositions oscillate on scales of 10 to 30 μm (Fig. 1). The greatest compositional variations (factor of 2 to 3) are in elements that can be divided into two groups that covary inversely: One group contains Mg, La, Ce, Ca, Ti, and Al, and the other Mn, Y, Sm, Nd, Th, Pr, and Fe (table S2). The zoning cannot reflect growth in a boundary layer that was depleted or enriched in allanite-compatible elements because melt trapped in the growing allanites lack such depletions or enrichments (*18*).

Ratios between the concentrations of chemically similar elements that substitute into the same crystallographic site in allanite, such as between the light and middle rare earth elements, can mirror compositional changes in melt from which the allanite grew in the same way as, for example, the Fe/Mg ratio in olivine mirrors that of the melt from which it grew (*19*). Two ratios that vary by a factor of ~2 in the allanites are MnO/MgO and La/Nd (Fig. 1). Each traces a distinct component of fractionation in rhyolitic magmas (*20*) and is essentially not fractionated by kinetic effects or coupled substitution in allanite because elements within each pair are similarly sized and charged and fit in the same crystallographic site. The effect of increasing fractionation on the composition of allanite is to progressively lower La/Nd and increase MnO/MgO in response to concomitant changes in the host melt (Fig. 2).

The variations in La/Nd and MnO/MgO in YTT allanites correlate smoothly (Fig. 2). In some crystals, zoning is normal (trending to lower La/Nd and higher MnO/MgO) or reverse, but in more than two-thirds of them, it is oscillatory, with a near-rim trend to a similar, more evolved composition (Fig. 1). Even though these allanites are present in one of the most evolved YTT pumices, some compositions match those for representative allanites from the least evolved YTT rhyolite (Fig. 2). Reversals to less evolved compositions (lower MnO/MgO and higher La/Nd) are typically abrupt and correspond to irregular boundaries marking zones with contrasting tone in backscattered electron (BSE) images.

La/Nd and MnO/MgO exchange coefficients enable us to predict how the YTT allanite compositions are related to fractionation of their parental melts (Fig. 2). Estimated La/Nd and MnO/MgO ratios for the rhyolitic melts overlap those for erupted glasses reported by Chesner (*15*), and the variation can be related by ~45% fractional crystallization (Fig. 2). This is comparable to an estimate of 40 to 50% fractionation based on the major element variability of YTT pumices (*15*). Affinity between allanite compositions and YTT melts is further suggested by the agreement between measured and predicted melt concentrations for elements such as Mn and Mg (fig. S1).

The young age of the allanites shows that the YTT eruption tapped a rhyolitic magma produced after the demise of the preceding (Middle Toba Tuff, 500 ka) caldera magma chamber. In addition, the continuity of the growth record shows that the antiquity of the allanites is not due solely to preferential preservation and/or recycling of crystals from older intrusions. The ~35,000 to 150,000 years of magmatic evolution recorded by individual allanites is comparable to crystallization intervals estimated largely by zircon dating (*21*) of the other voluminous (>1000 km^3) silicic magmas. Within this time frame, voluminous crystal-rich magmas can undergo fractionation differentiation by compaction and/or hindered crystal settling or can be thermally rejuvenated by an influx of mafic magma into the base of the subvolcanic reservoir (*22*, *23*).

The oscillatory zoning and disparate histories of the allanite (Fig. 1) require heterogeneous conditions of crystallization, whether in mush (>40% crystals) or liquid-rich domains. The irregular boundaries and compositional reversals in the allanites are suggestive of episodic dissolution due to mixing with hotter, less evolved magmas. Crystals may have been cycled between distinct batches of magma in Toba's reservoir by differential movement along boundaries between convectional zones (*24*, *25*) or mingling between recharge and resident magmas and/or during intrareservoir self-mixing (*26*, *27*). Although the range in melt variation required by the allanite zoning does not require some of the crystallization to have occurred in magma mush,

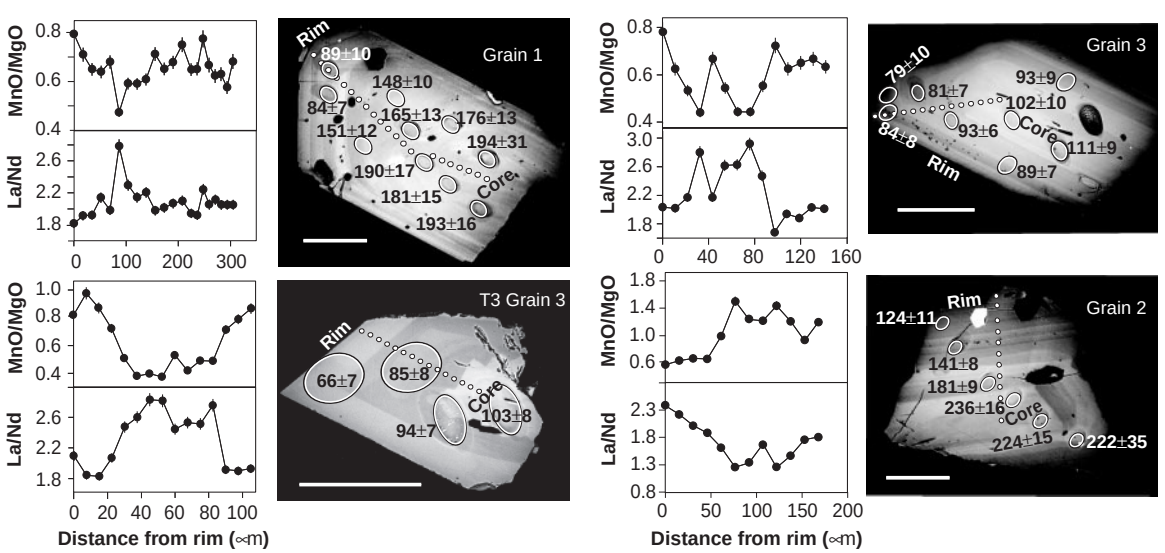


Fig. 1. Compositional and age results for YTT allanites illustrated by (i) ^{238}U-^{230}Th age distributions superimposed on BSE images of representative YTT allanites and (ii) core-to-rim compositional variations in MnO/MgO and La/Nd (note the different compositional scale for grain 2). Locations of in situ analyses are shown by open circles (ion probe) and white circles (electron probe). The contrast in BSE images represents differences in the abundances of high-mass elements, such as U, Th, and rare earth elements, and correlates with compositional changes measured by electron probe. Uncertainties are 1σ. The scale bar in BSE images is 100 μm.

it does not preclude it either, in which case the mush must have been periodically invaded by new silicic magma [compare (*28*)]. Evidently, magmatic conditions were frequently disrupted as the crystals grew, and the ages imply that greater disruption was closer to the time of eruption.

From ~225 to 110 ka, the allanite compositions are relatively restricted, excluding a single grain with evolved compositions (Fig. 3). Between ~110 and 75 ka, the compositional variability of the allanites is high. Only close to their rims do the allanite compositions converge (Figs. 1 and 3). On the basis of these patterns, we suggest that initially, and for a protracted interval of time, much of the Youngest Toba Tuff reservoir was relatively homogeneous, with melt compositions varying by <15% fractionation (or ~74 to 77 weight % SiO_2). The reservoir was nearly thermally stagnant, reflecting a heat balance perched between magmatic influxes and cooling of the system. Nearer to eruption, the diversity of melts sampled by the crystals increased substantially (melts related by up to 45% fractionation or ~70 to 77 weight % SiO_2). Mixing was probably less efficient as the system grew in size and diverse conditions of magma storage developed, resulting in domains of variably fractionated magmas and compositional zoning of the reservoir. Zoning of the magma that finally erupts (68 to 77 weight % SiO_2) could have developed even closer to eruption if the crystal-chemical variations arose in a mush rather than a liquid-dominated reservoir. Different magma batches may have intermittently coalesced in response to mass and heat input from the influx of new magma, as documented for the recent eruptions of Soufriere Hills and Ruapehu volcanoes (*8*, *29*), or when melts were expelled by gravitational collapse of critically thickened batches of magma mush (*22*). The likely voluminous domains of not-yet rigid magma mush probably enhanced the likelihood of cumulate crystals being reentrained (*12*), but those rare crystals with distinct compositions were probably harvested from relatively isolated batches of magma that were even closer to solidification. Final merging of magma in the Toba reservoir, rather than the periodic recharge that sustained the magmatic system for >100,000 years, could have catalyzed the cataclysmic eruption.

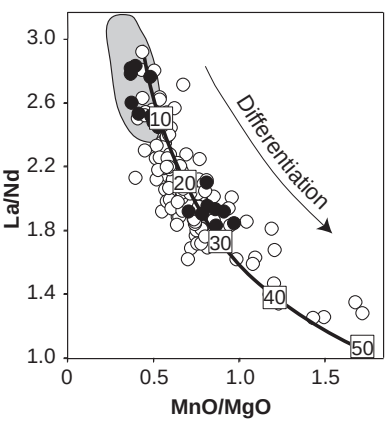


Fig. 2. Plot of La/Nd versus MnO/MgO for allanites from evolved YTT rhyolite. Curve shows allanite compositions expected for crystallization from melts related by fractionating the quartz-plagioclase-sanidine-biotite phenocryst assemblage in modal proportions observed by (*15*), beginning from the least evolved melt composition. A La/Nd exchange coefficient, $D^{La/Nd}$, relating allanite composition to melt composition is 1.7 ± 0.1 [1 SD; computed from data of (*38*, *39*)] and agrees well with values of 1.5 to 1.6 calculated by applying the model of Blundy and Wood (*40*) to the structural data of Dollase (*41*). $D^{La/Nd}$ is essentially constant over much of the range of low-to-high silica rhyolites, even though absolute partition coefficients increase. A $D^{MnO/MgO}$ value of 1.4 ± 0.3 (1 SD) based on data for high-silica rhyolites [data of (*39*, *42*)] is less constrained because of a smaller number of partitioning data, but it agrees with the value of 1.4 based on the Blundy and Wood (*40*) model. Compositions in single grains (e.g., grain 3, black circles) may overlap nearly the entire range of observed allanite compositions, including allanites from the least evolved (68 weight % SiO_2) rhyolite (gray field) reported by (*43*).

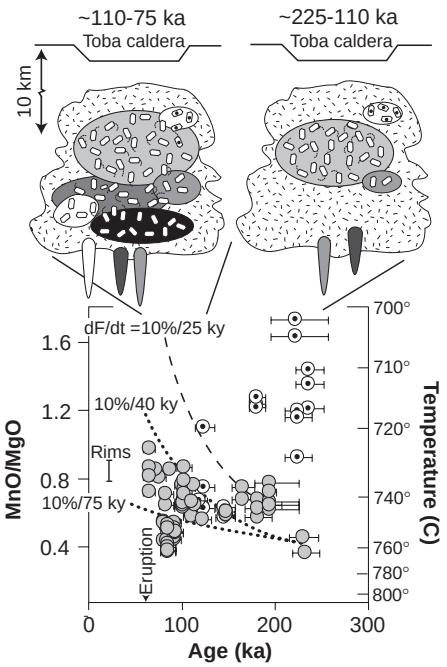


Fig. 3. Temporal variation in MnO/MgO for YTT allanites and cartoons depicting the magmachronology of rhyolite beneath Toba caldera. The depth of the system is from (*15*). Temperatures of magma during allanite growth are based on covariation of temperature and MnO/MgO reported by (*44*) for experimental crystallization of YTT magma compositions at 100 to 200 MPa. Reference to dashed curves for different rates of MnO/MgO fractionation (*dF/dt* = rate of crystal-liquid fractionation, starting at two different compositions recorded in the oldest allanite cores) based on the mineralogy the YTT emphasizes the divergence of the data from simple evolutionary trends. Most allanite compositions (gray circles) between ~225 and 110 ka are restricted and reflect a body of rhyolitic magma that is relatively invariant compositionally. Rare grains (e.g., grain 2) reflect isolated batches of highly evolved magmas (white circles with black dots). The increase in diversity of allanite composition between 110 and 75 ka is produced by the interaction and mingling of differentially fractionated batches of rhyolitic magma with temperatures between ~760° and 715°C. Final mixing (not shown) gathers crystals into the most evolved batch of rhyolite in the reservoir that then erupts at 75 ka.

Our results demonstrate that the components of a huge subvolcanic magma reservoir may unite crystals that probe magmatic evolution in space and time, and that intrusions of silicic magma may undergo a transition between homogeneous and heterogeneous states during their storage in Earth's crust. A corollary is that in chemically and/or isotopically zoned bodies of magma (*10*, *27*), different crystal-zoning profiles may reflect spatial as well as temporal variations in the magma reservoir. Our results predict that the crystal-rich residue remaining after eruption of the YTT magma would form a compositionally zoned pluton that is locally monotonous but complexly zoned at a mineral scale, features that are increasingly observed in the plutonic record (*12*, *30*). Because the YTT magma reservoir grew by piecemeal accumulation [compare (*6*, *7*)] with mingling between successive additions of magma, crystals from domains of the reservoir that did not erupt, such as any cumulate pile underpinning the more liquid portions of the reservoir (*31*, *32*), might record this evolution as well. Generation of the YTT magma by melting and remobilization of a young granitic pluton (*33*) is unlikely because the amount of crystallization recorded by the allanites is so much less than expected for solidification of an intrusion. Instead, the YTT magma accumulated and evolved over a period of >100,000 years.

References and Notes

1. C. A. Chesner, W. I. Rose, A. Deino, R. Drake, J. A. Westgate, *Geology* **19**, 200 (1991).
2. M. R. Rampino, S. Self, *Nature* **359**, 50 (1992).
3. M. R. Rampino, S. Self, *Science* **262**, 1955 (1993).
4. H. M. Iyer in, *Volcanic Seismology*, P. Gasparini, R. Scarpa, K. Aki, Eds. (Springer-Verlag, Berlin, 1992), pp. 299–338.
5. A. F. Glazner, J. M. Bartley, D. S. Coleman, W. Gray, R. Z. Taylor, *GSA Today* **14**, 4 (2004).
6. R. B. Hanson, A. F. Glazner, *Geology* **23**, 213 (1995).
7. A. S. Yoshinobu, D. A. Okaya, S. R. Paterson, *J. Struct. Geol.* **20**, 1205 (1998).
8. M. D. Murphy, R. S. J. Sparks, J. Barclay, M. R. Carroll, T. S. Brewer, *J. Petrol.* **41**, 21 (2000).
9. S. Blake, *Nature* **289**, 783 (1981).
10. J. P. Davidson, F. J. T. Tepley, *Science* **275**, 826 (1997).
11. G. S. Wallace, G. W. Bergantz, *Earth Planet. Sci. Lett.* **202**, 133 (2002).
12. J. Blundy, N. Shimizu, *Earth Planet. Sci. Lett.* **102**, 178 (1991).

13. A. T. Anderson, A. M. Davis, F. Lu, *J. Petrol.* **41**, 449 (2000).
14. M. R. Reid, C. D. Coath, T. M. Harrison, K. D. McKeegan, *Earth Planet. Sci. Lett.* **150**, 27 (1997).
15. C. A. Chesner, *J. Petrol.* **39**, 397 (1998).
16. Materials and analytical methods are available as supporting material on *Science* Online.
17. Model ^{238}U-^{230}Th ages are derived as described in (*14*). See (*34*) for a review of ^{230}Th dating in magmatic systems. The reported ages are taken to be those of crystallization because of the relatively tight packing of ions within allanite and the tetravalent charge of Th [compare (*35*)]. Reequilibration of Th during magmatic residence will be insignificant: Based on the predictive model of Fortier and Giletti (*36*), Th diffusion would only affect ~2 μm in allanite over a period of 100,000 years at the highest reported temperature of the YTT magma (~780°C). Uncertainty in the calculated ages due to possible variation of initial ^{230}Th/^{232}Th during magmatic evolution can be evaluated by assuming that observed eruption-age ^{230}Th/^{232}Th activity ratio variations of the YTT (0.358 to 0.433) are representative of the initial range of Th-isotope composition. For reported ages <120 ka, the uncertainty in age associated with the initial ratio is within the analytical uncertainty on the ages, except for the youngest allanite domains which could not have grown from melts that had Th isotope compositions significantly different from that of their host. Reported ages that are 120 to 200 ka could be at most a few percent to, for the older of these, as much as 27% older than allowed by the analytical uncertainty. Those few allanites with ages >200 ka could be substantially older. Thus, the age ranges reported here are conservative.
18. J. B. Thomas, R. J. Bodnar, N. Shimizu, C. Chesner, in *Zircon*, J. M. Hanchar, P. W. O. Hoskin, Eds. (Mineralogical Society of America, Washington, DC, 2004), vol. 53, chap. 3.
19. P. L. Roeder, R. F. Emslie, *Contrib. Mineral. Petrol.* **29**, 275 (1970).
20. MnO/MgO in residual melts of silicic magmas typically increases with fractionation of major mafic silicates. La/Nd also increases except when the fractionating assemblage includes sufficient quantities of allanite, chevkinite, and/or monazite that are rich in light rare earth elements (*37*).
21. M. R Reid, in *Treatise on Geochemistry*, H. D. Holland, K. K Turekian, Eds. (Elsevier, Amsterdam, 2003), vol. 3, chap. 3.05.
22. O. Bachmann, G. W. Bergantz, *J. Petrol.*, **45**, 1565 (2004).
23. O. Bachmann, G. W. Bergantz, *Geology* **27**, 447 (2003).
24. B. D. Marsh, M. R. Maxey, *J. Volcanol. Geotherm. Res.* **24**, 95 (1985).
25. V. R. Troll, H. U. Schmincke, *J. Petrol.* **43**, 243 (2002).
26. G. W. Bergantz, *J. Struct. Geol.* **22**, 1297 (2000).
27. S. Couch, R. S. J. Sparks, M. R. Carroll, *Nature* **411**, 1037 (2001).
28. S. Turner, R. George, D. A. Jerram, N. Carpenter, C. Hawkesworth, *Earth Planet. Sci. Lett.* **214**, 279 (2003).
29. M. Nakagawa, K. Wada, T. Thordarson, C. P. Wood, J. A. Gamble, *Bull. Volcanol.* **61**, 15 (1999).
30. D. M. Robinson, C. F. Miller, *Am. Mineral.* **84**, 1346 (1999).
31. C. A. Bachl, C. F. Miller, J. S. Miller, J. E. Faulds, *GSA Bull.* **113**, 1213 (2001).
32. T. H. Druitt, C. R. Bacon, *Trans. R. Soc. Edinburgh Earth Sci.* **79**, 289 (1988).
33. I. N. Bindeman, J. W. Valley, *Geology* **28**, 719 (2000).
34. M. Condomines, P.-J. Gauthier, O. Sigmarsson, in *Uranium-Series Geochemistry*, B. Bourdon, G. M. Henderson, C. C. Lundstrom, S. P. Turner, Eds. (Mineralogical Society of America, Washington, DC, 2003), vol. 52, chap. 4.
35. E. Dowty, *Am. Mineral.* **65**, 174 (1980).
36. S. M. Fortier, B. J. Giletti, *Science* **245**, 1481 (1989).
37. C. F. Miller, D. W. Mittlefehldt, *Geology* **10**, (1982).
38. C. K. Brooks, P. Henderson, J. G. Ronsbo, *Mineral. Mag.* **44**, 157 (1981).
39. G. A. Mahood, W. Hildreth, *Geochim. Cosmochim. Acta* **47**, 11 (1983).
40. J. Blundy, B. Wood, *Nature* **372**, 452 (1994).
41. W. A. Dollase, *Am. Mineral.* **56**, 447 (1971).
42. A. Ewart, W. L. Griffin, *Chem. Geol.* **117**, 251 (1994).
43. C. A. Chesner, A. D. Ettlinger, *Am. Mineral.* **74**, 750 (1989).
44. J. E. Gardner, P. W. Layer, M. J. Rutherford, *Geology* **30**, 347 (2002).

45. We are grateful to C. Chesner for samples; C. Coath, F. Ramos, and F. Kyte for analytical help; and especially J. Simon and G. Bergantz for insightful discussions. Anonymous referees provided very helpful reviews. Funded by NSF grants EAR-9706519 and EAR-0003601. The University of California, Los Angeles (UCLA), ion microprobe is partially subsidized by a grant from the NSF Instrumentations and Facilities Program.

Supporting Online Material
www.sciencemag.org/cgi/content/full/305/5686/991/DC1
Materials and Methods
Tables S1 and S2
Fig. S1
References

19 February 2004; accepted 9 July 2004

More Intense, More Frequent, and Longer Lasting Heat Waves in the 21st Century

Gerald A. Meehl* and Claudia Tebaldi

A global coupled climate model shows that there is a distinct geographic pattern to future changes in heat waves. Model results for areas of Europe and North America, associated with the severe heat waves in Chicago in 1995 and Paris in 2003, show that future heat waves in these areas will become more intense, more frequent, and longer lasting in the second half of the 21st century. Observations and the model show that present-day heat waves over Europe and North America coincide with a specific atmospheric circulation pattern that is intensified by ongoing increases in greenhouse gases, indicating that it will produce more severe heat waves in those regions in the future.

There is no universal definition of a heat wave, but such extreme events associated with particularly hot sustained temperatures have been known to produce notable impacts on human mortality, regional economies, and ecosystems (*1*–*3*). Two well-documented examples are the 1995 Chicago heat wave (*4*) and the Paris heat wave of 2003 (*5*). In each case, severe hot temperatures contributed to human mortality and caused widespread economic impacts, inconvenience, and discomfort.

In a future warmer climate with increased mean temperatures, it seems that heat waves would become more intense, longer lasting, and/or more frequent (*6*, *7*). However, analyses of future changes in other types of extreme events, such as frost days, show that changes are not evenly distributed in space but are characterized instead by particular patterns related to larger scale climate changes (*8*). Here, we examine future behavior of heat waves in a global coupled climate model, the Parallel Climate Model (PCM). This model has a latitude-longitude resolution of about 2.8° in the atmosphere and a latitude-longitude resolution of less than 1° in the ocean, and it contains interacting components of atmosphere, ocean, land surface, and sea ice. The PCM has been used extensively to simulate climate variability and climate change in a variety of applications for 20th- and 21st-century climate (*6*, *8*–*13*). We analyzed a four-member ensemble (i.e., the model was run four times from different initial states and the four members were averaged together to reduce noise) for 20th-century climate and a five-member ensemble for 21st-century climate. The former includes the major observed forcings for the 20th century encompassing greenhouse gases, sulfate aerosols, ozone, volcanic aerosols, and solar variability (*13*). The latter uses a "business-as-usual" scenario, which assumes little in the way of policy intervention to mitigate greenhouse gas emissions in the 21st century (*14*). We define the present-day reference period as 1961 to 1990 for model and observations and the future as the time period from 2080 to 2099.

First, we sought to define a heat wave. Many definitions could apply to heat waves that quantify the duration and/or intensity of either nighttime minima or daytime maxima (*4*, *5*, *15*, *16*). Here, we used two definitions of heat waves; each has been shown to be associated with substantial societal impacts on human health and economies. The first (*4*) evolved from a study of the 1995 Chicago heat wave; it concentrates on the severity of an annual "worst heat event" and suggests that several consecutive nights with no relief from very warm nighttime minimum temperatures may be most important for health impacts. For present-day climate for North America and Europe (Fig. 1), the means of three consecutive warmest nights for observations and the model show good agreement. Heat waves presently are more severe in the southeast United States (large areas greater than 24°C) and less severe in the northwest United States (equally large areas less than 16°C; Fig. 1, A and C). For Europe, there is more of a north-south gradient in both obser-

National Center for Atmospheric Research (NCAR), Post Office Box 3000, Boulder, CO 80307, USA.

*To whom correspondence should be addressed. E-mail: meehl@ncar.ucar.edu

vations and the model (Fig. 1, B and D), with more severe heat waves in the Mediterranean region (most countries bordering the Mediterranean have values greater than 20°C) and less severe heat in northern Europe (many areas less than 16°C).

Future changes of worst 3-day heat waves defined in this way in the model are not uniformly distributed in space but instead show a distinct geographical pattern (Fig. 1, E and F). Though differences are positive in all areas, indicative of the general increase of nighttime minima, heat wave severity increases more in the western and southern United States and in the Mediterranean region, with heat wave severity showing positive anomalies greater than 3°C in those regions. Thus, many of the areas most susceptible to heat waves in the present climate (greatest heat wave severity in Fig. 1, A to D) experience the greatest increase in heat wave severity in the future. But other areas not currently as susceptible, such as northwest North America, France, Germany, and the Balkans, also experience increased heat wave severity in the 21st century in the model.

The second way we chose to define a heat wave is based on the concept of exceeding specific thresholds, thus allowing analyses of heat wave duration and frequency. Three criteria were used to define heat waves in this way, which relied on two location-specific thresholds for maximum temperatures. Threshold 1 (T1) was defined as the 97.5th percentile of the distribution of maximum temperatures in the observations and in the simulated present-day climate (seasonal climatology at the given location), and T2 was defined as the 81st percentile. A heat wave was then defined as the longest period of consecutive days satisfying the following three conditions: (i) The daily maximum temperature must be above T1 for at least 3 days, (ii) the average daily maximum temperature must be above T1 for the entire period, and (iii) the daily maximum temperature must be above T2 for every day of the entire period (*16*).

Because the Chicago heat wave of 1995 and the Paris heat wave of 2003 had particularly severe impacts, we chose grid points from the model that were close to those two locations to illustrate heat wave characteristics. This choice was subjective and illustrative given that there are, of course, other well-known heat waves from other locations. Also, we are not suggesting that a model grid point is similar to a particular weather station; we picked these grid points because they represent heat wave conditions for regions representative of Illinois and France in the model, and therefore they can help identify processes that contribute to changes in heat waves in the future climate in those regions. We chose comparable grid points from the National Centers for Environmental Prediction (NCEP)/NCAR reanalyses that used assimilated observational data (*17*, *18*) for comparison to the model results.

For both the Paris- and Chicago-area grid points for the five-ensemble members, a future increase in heat wave occurrence is predicted (Fig. 2, A and B). For Chicago, the number of present-day heat waves (1961 to 1990) ranges from 1.09 to 2.14 heat waves per year for the four-ensemble members, whereas for future climate, the range shifts to between 1.65 and 2.44. Thus, the ensemble mean heat wave occurrence increases 25% from 1.66 to 2.08 heat waves per year. The current observed value from NCEP for 1961 to 1990 lies within the present-day range from the model with a value of 1.40 events per year. For Paris, the model ranges from 1.18 to 2.17 heat waves per year at present (the value from the NCEP reanalysis lies barely outside this range at 1.10 days), with the future range shifting to between 1.70 and 2.38. Thus, the ensemble mean heat wave occurrence for the Paris grid point increases 31% from 1.64 to 2.15 heat waves per year. Both observed values from NCEP fall well short of the future range from the model, indicative of the shift to more heat waves per year in the future climate.

There is a corresponding increase in duration at both locations (Fig. 2, C and D). For Chicago, present-day average duration of heat waves from the four–model ensemble members ranges from 5.39 to 8.85 days, encompassing the observed value from NCEP at 6.29 days. For Paris, the present-day model range is 8.33 to 12.69 days, with the NCEP observation lying within that range at 8.40 days. For future climate at the Paris grid point, there is a shift to longer lived heat waves with average duration increasing from 11.39 days to 17.04 days. For both of these regions, similar to what was found for the number of heat waves, the corresponding grid point values from the NCEP reanalyses show the duration to be within or very near the range of the present-day model ensemble members but not the future ensemble members, indicative of the

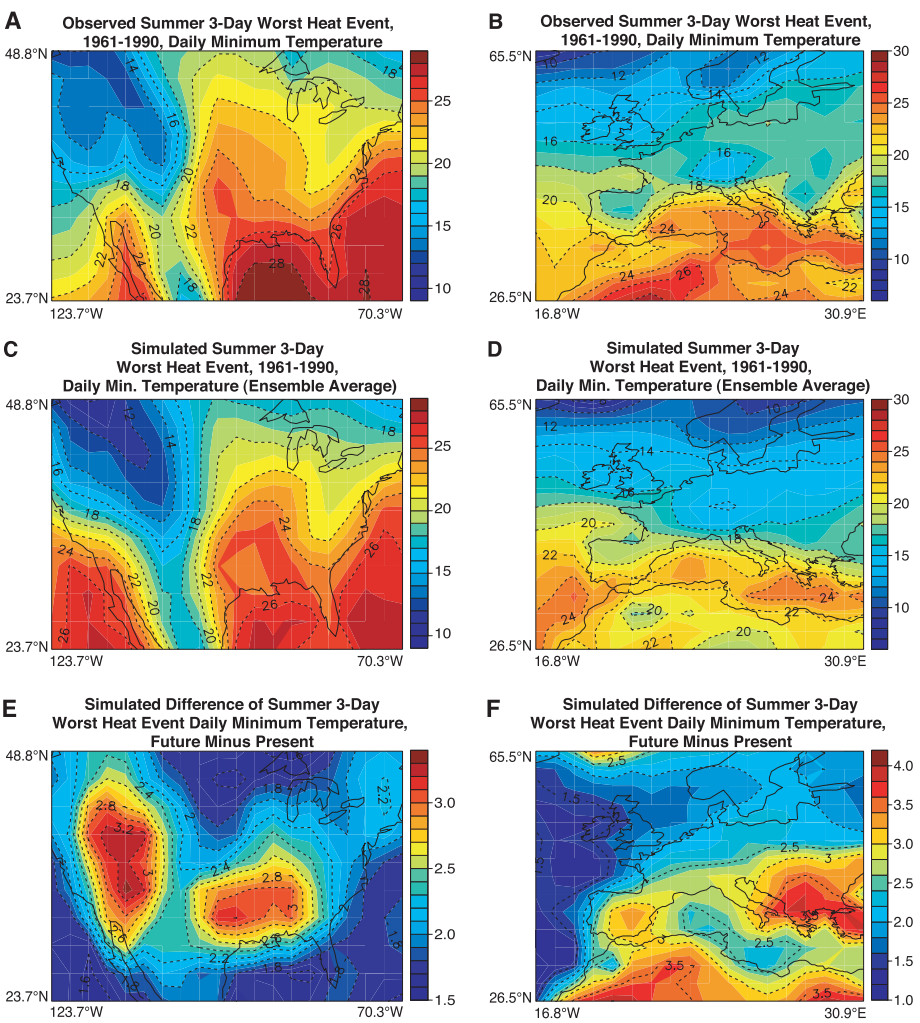


Fig. 1. Heat wave severity as the mean annual 3-day worst (warmest) nighttime minima event (*4*) from NCEP/NCAR reanalyses, 1961 to 1990, for North America (°C) (**A**) and Europe (**B**), and from the model for North America (**C**) and Europe (**D**). The changes of 3-day worst (warmest) nighttime minima event from the model, future (2080 to 2099) minus present (1961 to 1990) for North America (°C) (**E**) and Europe (**F**) are also shown.

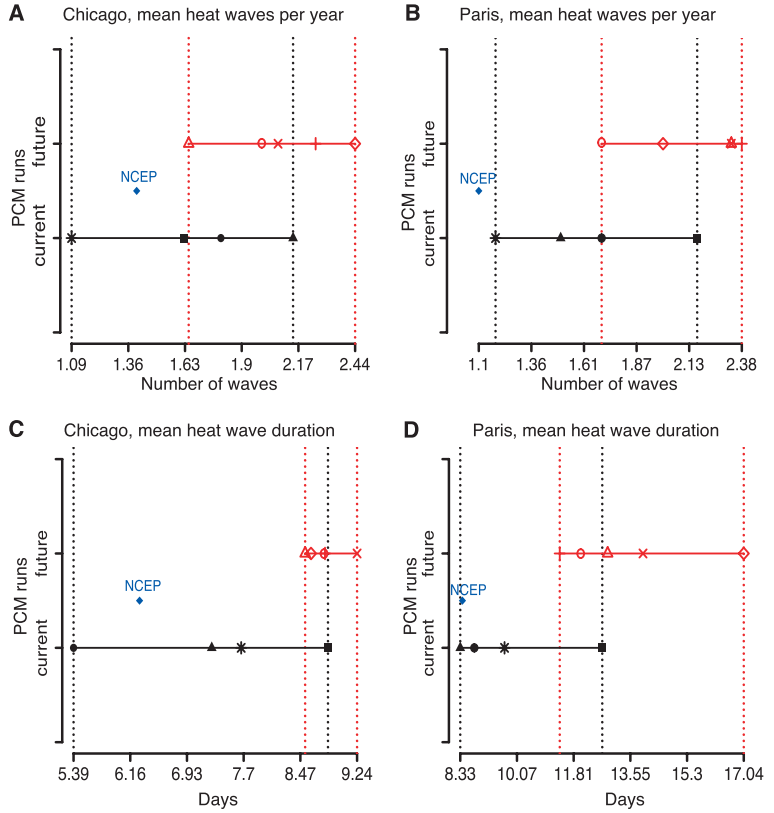


Fig. 2. Based on the threshold definition of heat wave (*16*), mean number of heat waves per year near Chicago (**A**) and Paris (**B**) and mean heat wave duration near Chicago (**C**) and Paris (**D**) are shown. In each panel, the blue diamond marked NCEP indicates the value computed from NCEP/NCAR reanalysis data. The black segment indicates the range of values obtained from the four ensemble members of the present-day (1961 to 1990) model simulation. The red segment indicates the range of values obtained from the five ensemble members of the future (2080 to 2099) model simulation. The single members are marked by individual symbols along the segments. Dotted vertical lines facilitate comparisons of the simulated ranges/observed value.

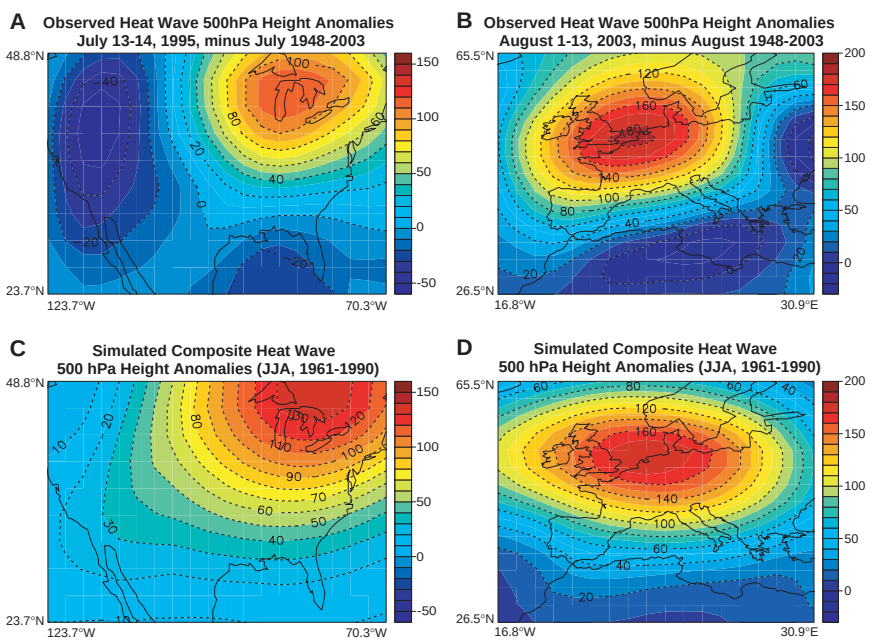


Fig. 3. Height anomalies at 500 hPa (gpm) for the 1995 Chicago heat wave (anomalies for 13 to 14 July 1995 from July 1948 to 2003 as base period), from NCEP/NCAR reanalysis data (**A**) and the 2003 Paris heat wave (anomalies for 1 to 13 August 2003 from August 1948 to 2003 as base period), from NCEP/NCAR reanalysis data (**B**). Also shown are anomalies for events that satisfy the heat wave criteria in the model in present-day climate (1961 to 1990), computed at grid points near Chicago (**C**) and Paris (**D**). In both cases, the base period is summer [June, July, August (JJA)], 1961 to 1990.

shift in the model to more and longer lived heat waves in future climate.

Heat waves are generally associated with specific atmospheric circulation patterns represented by semistationary 500-hPa positive height anomalies that dynamically produce subsidence, clear skies, light winds, warm-air advection, and prolonged hot conditions at the surface (*15*, *19*). This was the case for the 1995 Chicago heat wave and 2003 Paris heat wave (Fig. 3, A and B), for which 500-hPa height anomalies of over +120 geopotential meters (gpm) over Lake Michigan for 13 to 14 July 1995 and +180 gpm over northern France for 1 to 13 August 2003 are significant at greater than the 5% level according to a Student's *t* test. A stratification based on composite present-day heat waves from the model for these two locations over the period of 1961 to 1990 (Fig. 3, C and D) shows comparable amplitudes and patterns, with positive 500-hPa height anomalies in both regions greater than +120 gpm and significance exceeding the 5% level for anomalies of that magnitude.

There is an amplification of the positive 500-hPa height anomalies associated with a given heat wave for Chicago and Paris for future minus present climate (Fig. 4, A and B). Statistically significant (at greater than the 5% level) ensemble mean heat wave 500-hPa differences for Chicago and Paris in the future climate compared with present-day are larger by about 20 gpm in the model (comparing Fig. 4, A and B, with Fig. 3, C and D).

The future modification of heat wave characteristics with a distinct geographical pattern (Fig. 1, E and F) suggests that a change in climate base state from increasing greenhouse gases could influence the pattern of those changes. The mean base state change for future climate shows 500-hPa height anomalies of nearly +55 gpm over the upper Midwest, and about +50 gpm over France for the end of the 21st century (Fig. 4, C and D, all significant above the 5% level). The 500-hPa height increases over the Mediterranean and western and southern United States for future climate are directly associated with more intense heat waves in those regions (Fig. 1, E and F), thus confirming the link between the pattern of increased 500-hPa heights for future minus present-day climate and increased heat wave intensity in the future climate. A comparable pattern is present in an ensemble of seven additional models for North America for future minus present-day climate, with somewhat less agreement over Europe (fig. S1). In that region, there is still the general character of largest positive anomalies over the Mediterranean and southern Europe regions, and smaller positive anomalies to the north (fig. S1), but largest positive values occur near Spain, as opposed to the region near Greece as in our model (Fig. 4D). This also corresponds to a similar pattern for increased standard deviations of both summertime nighttime minimum and daytime

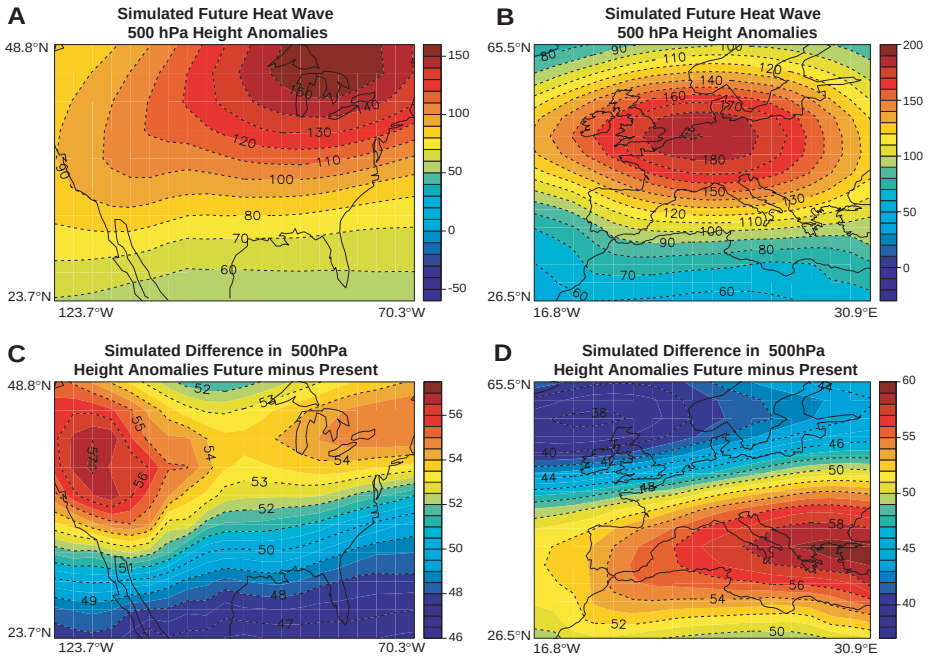


Fig. 4. Height anomalies at 500 hPa (gpm) for events that satisfy the heat wave criteria in the model in future climate (2080 to 2099) for grid points near Chicago (**A**) and Paris (**B**), using the same base period as in Fig. 3, C and D. Also shown are changes (future minus present) in the model's 500-hPa height mean base state, for North America (**C**) and Europe (**D**).

maximum temperatures (fig. S2). This is consistent with a widening of the distribution of temperatures in addition to a shift in the mean (*5*), and suggests that there is an increase in heat wave occurrence beyond that driven by changes in the mean circulation.

The 500-hPa height anomalies are most strongly related to positive warm season precipitation anomalies over the Indian monsoon region and associated positive convective heating anomalies that drive mid-latitude teleconnection patterns (such as those in Fig. 4, C and D) in response to anomalous tropical convective heating in future climate (figs. S3 to S5). Thus, areas already experiencing strong heat waves (e.g., southwest, midwest, and southeast United States and the Mediterranean region) could experience even more intense heat waves in the future. But other areas (e.g., northwest United States, France, Germany, and the Balkans) could see increases of heat wave intensity that could have more serious impacts because these areas are not currently as well adapted to heat waves.

References and Notes

1. C. Parmesan *et al.*, *Bull. Am. Meteorol. Soc.* **81**, 443 (2000).
2. D. R. Easterling *et al.*, *Science* **289**, 2068 (2000).
3. World Health Organization (WHO), "The health impacts of 2003 summer heat waves," WHO Briefing Note for the Delegations of the 53rd session of the WHO Regional Committee for Europe, Vienna, Austria, 8 to 11 September 2003; available at www.euro.who.int/document/Gch/HEAT-WAVES%20RC3.pdf.
4. T. R. Karl *et al.*, *Bull. Am. Meteorol. Soc.* **78**, 1107 (1997).
5. C. Schar *et al.*, *Nature* **427**, 332 (2004).
6. U. Cubasch *et al.*, in *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, J. T. Houghton *et al.*, Eds. (Cambridge Univ. Press, Cambridge, 2001), pp. 525–582.
7. T. R. Karl, K. E. Trenberth, *Science* **302**, 1719 (2003).
8. G. A. Meehl *et al.*, *Clim. Dyn.*, in press.
9. W. M. Washington *et al.*, *Clim. Dyn.* **16**, 755 (2000).
10. A. Dai *et al.*, *Geophys. Res. Lett.* **28**, 4511 (2001).
11. G. A. Meehl *et al.*, *J. Clim.* **16**, 426 (2003).
12. B. D. Santer *et al.*, *Science* **301**, 479 (2003).
13. G. A. Meehl *et al.*, *J. Clim.*, in press.
14. A. Dai *et al.*, *J. Climate* **14**, 485 (2001) describes the business-as-usual scenario as similar to the A1B emissions scenario of the Intergovernmental Panel on Climate Change (IPCC) Special Report on Emission Scenarios (SRES) (*20*), with CO_2 rising to about 710 parts per million by volume by 2100, SO_2 emissions declining to less than half the present value by 2100, CH_4 stabilized at 2500 parts per billion by volume in 2100, N_2O as in the IPCC IS92 emissions scenario (*21*), and halocarbons following a preliminary version of the SRES A1B scenario.
15. M. A. Palecki *et al.*, *Bull. Am. Meteorol. Soc.* **82**, 1353 (2001).
16. R. Huth *et al.*, *Clim. Change* **46**, 29 (2000).
17. E. Kalnay *et al.*, *Bull. Am. Meteorol. Soc.* **77**, 437 (1996).
18. Archived observations from surface weather stations, weather balloons, satellites, and other sources are interpolated to a regular grid in a weather forecast model, and the model is run at regular intervals over past time periods to produce a dynamically consistent time-evolving representation of the observed historical climate state.
19. K. E. Kunkel *et al.*, *Bull. Am. Meteorol. Soc.* **77**, 1507 (1996).
20. N. Nakicenovic *et al.*, *IPCC Special Report on Emission Scenarios* (Cambridge Univ. Press, Cambridge, 2000).
21. J. Legget *et al.*, *Climate Change 1992: The Supplementary Report to the IPCC Scientific Assessment* (Cambridge Univ. Press, New York, 1992), pp. 69–75.

22. We thank D. Nychka for discussions; G. Branstator for the convective heating anomaly results; and L. Buja, J. Arblaster, and G. Strand for assistance on the CMIP2+ results from the Coupled Model Intercomparison Project, phase 2 plus (www-pcmdi.llrl.gov/cmip). This work was supported in part by the Weather and Climate Impact Assessment Initiative at the National Center for Atmospheric Research. A portion of this study was also supported by the Office of Biological and Environmental Research, U.S. Department of Energy, as part of its Climate Change Prediction Program, and the National Center for Atmospheric Research. The National Center for Atmospheric Research is sponsored by NSF.

Supporting Online Material
www.sciencemag.org/cgi/content/full/305/5686/994/DC1
Figs. S1 to S5
References

2 April 2004; accepted 9 July 2004

Discovery of Symbiotic Nitrogen-Fixing Cyanobacteria in Corals

Michael P. Lesser,[1]* Charles H. Mazel,[2] Maxim Y. Gorbunov,[3] Paul G. Falkowski[3,4]

Colonies of the Caribbean coral *Montastraea cavernosa* exhibit a solar-stimulated orange-red fluorescence that is spectrally similar to a variety of fluorescent proteins expressed by corals. The source of this fluorescence is phycoerythrin in unicellular, nonheterocystis, symbiotic cyanobacteria within the host cells of the coral. The cyanobacteria coexist with the symbiotic dinoflagellates (zooxanthellae) of the coral and express the nitrogen-fixing enzyme nitrogenase. The presence of this prokaryotic symbiont in a nitrogen-limited zooxanthellate coral suggests that nitrogen fixation may be an important source of this limiting element for the symbiotic association.

The success of scleractinian corals since the Triassic (*1*) has been attributed to the establishment of a mutualistic symbiosis between the cnidarian host and a diverse group of endosymbiotic dinoflagellates (zooxanthellae). Zooxanthellae, which are localized within gastrodermal cells of the cnidarian host, can provide more than

100% of the carbon requirements of the animal partner, primarily in the form of carbohydrates and low–molecular-weight lipids. Experimental manipulations of zooxanthellate corals suggest that inorganic nitrogen limits the growth and abundance of zooxanthellae in the coral; indeed, this limitation has been suggested to be essential for the stability of the symbiotic association (*2*, *3*).

In addition to zooxanthellae, a variety of bacteria appear to be associated with scleractinian corals (*4*–*6*), and although these associations also appear to be widely distributed, stable, and nonpathogenic, the function of these bacteria remains largely unknown. However, the presence of cyanobacteria is associated with photosynthesis-dependent nitrogen fixation on coral reefs (*7*) and is suggested to be responsible for nitrogen fixation in living coral tissue (*8*). In this paper, we show that large numbers of endosymbiotic cyanobacteria capable of fixing nitrogen occur in a common scleractinian coral, *Montastraea cavernosa.*

Scleractinian corals, including *M. cavernosa* (*9*–*11*), express a variety of fluorescent proteins, but colonies of *M. cavernosa* have also been observed to fluoresce orange during the daytime (Fig. 1A). This fluorescence is not due to a fluorescent protein but to phycoerythrin. In vivo excitation/emission spectra of these corals showed an emission peak at 580 nm and a shoulder at 630 nm, with excitation bands at 505 and 571 nm (Fig. 1B). Although this spectral signature is similar to those reported for red fluorescent proteins of corals (*12*, *13*), the two excitation peaks also corresponded to absorption by phycoerythrin in marine cyanobacteria that contain both the phycourobilin and phycoerythrobilin chromphores (*14*, *15*). Immunoblots (*16*) of coral homogenates challenged with a polyclonal antibody against phycoerythrin revealed a positive cross-reaction with the 18- to 20-kD β-polypeptide of phycoerythrin (Fig. 1C). These results clearly suggest that intracellular cyanobacteria are associated with the coral.

Fluorescence lifetime analyses (*16*) indicate that the 580-nm excited state is dominated by a single component with a 3.93-ns time constant (Fig. 1D). The slow single-exponential decay of the orange pigment is longer than described for fluorescent proteins (2.6 to 3.7 ns) (*11*, *17*, *18*) and suggests energetic isolation and the absence of excitation energy transfer out of the chromophore. In contrast, phycoerythrin fluorescence normally observed in cyanobacteria exhibits faster kinetics in vivo owing to efficient energy transfer within the phycobilisomes. The lifetime data and the daytime fluorescence indicate that the energy coupling of these pigments to primary photochemistry in the symbiotic cyanobacteria is weak, leading to the relatively high quantum yield of fluorescence for this pigment.

Epifluorescence microscopy (*16*) of host tissue homogenates revealed zooxanthellae exhibiting red chlorophyll fluorescence, as

[1]Department of Zoology and Center for Marine Biology, University of New Hampshire, Durham, NH 03824, USA. [2]Physical Sciences, 20 New England Business Center, Andover, MA 01810, USA. [3]Environmental Biophysics and Molecular Ecology Program, Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, NJ 08901, USA. [4]Department of Geological Sciences, Rutgers University, Piscataway, NJ, USA.

*To whom correspondence should be addressed. E-mail: mpl@cisunix.unh.edu

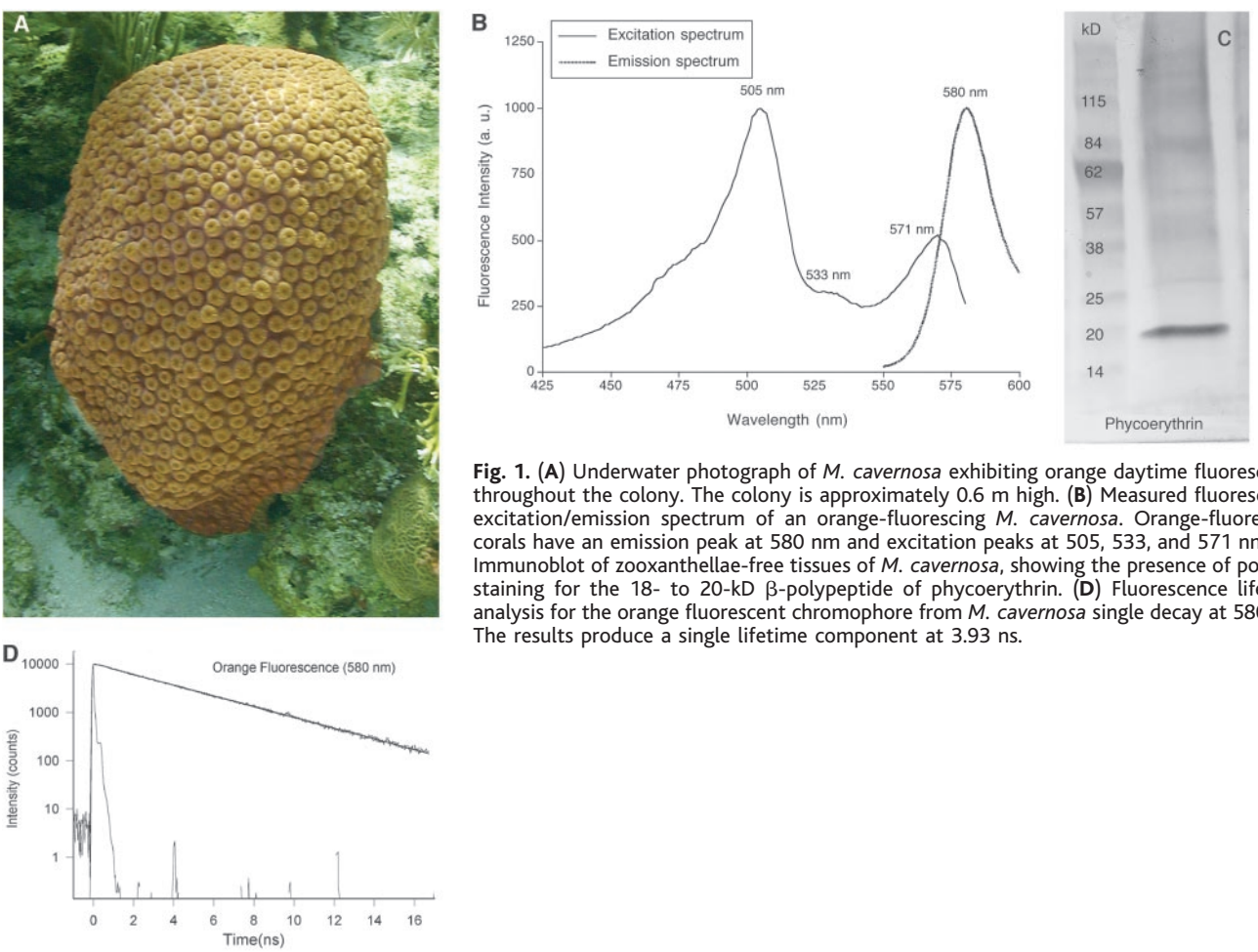


Fig. 1. (**A**) Underwater photograph of *M. cavernosa* exhibiting orange daytime fluorescence throughout the colony. The colony is approximately 0.6 m high. (**B**) Measured fluorescence excitation/emission spectrum of an orange-fluorescing *M. cavernosa*. Orange-fluorescing corals have an emission peak at 580 nm and excitation peaks at 505, 533, and 571 nm. (**C**) Immunoblot of zooxanthellae-free tissues of *M. cavernosa*, showing the presence of positive staining for the 18- to 20-kD β-polypeptide of phycoerythrin. (**D**) Fluorescence lifetime analysis for the orange fluorescent chromophore from *M. cavernosa* single decay at 580 nm. The results produce a single lifetime component at 3.93 ns.

well as many smaller orange-fluorescent cells resembling cyanobacteria (Fig. 2A). An analysis of tissue homogenates using flow cytometry (*16*) showed a distinct phycoerythrin signature and a size range for these cells of 1.0 to 3.0 μm in diameter. The number of phycoerythrin-positive cells from fluorescent samples of *M. cavernosa*, normalized to surface area, ranged from 1.14×10^7 to 2.55×10^7 cells cm^{-2}, whereas nonfluorescent colonies have $\ll 10^2$ phycoerythrin-positive cells per square centimeter. Transmission electron micrographs (*16*) of the coral tissue revealed that the cyanobacteria-like cells are located in the epithelial cells of the animal host and are surrounded by host membrane (Fig. 2B). The cyanobacteria-like cells exhibit an unusual arrangement of their thylakoid membranes that cross randomly throughout the cell and occasionally appear both expanded and appressed. These cells also appear to have fewer electron-dense phycobilisomes associated with the thylakoid membranes than has been reported for other cyanobacteria. Immunogold probing of thin sections (Fig. 2C), using the antibody to phycoerythrin, showed that the antibody binds significantly [analysis of variance (ANOVA), $P < 0.001$] to the cyanobacterial cells [134.9 particles per cell ± 19.4 (SE)] when compared to controls [6.5 particles per cell ± 2.1 (SE)] without the primary antibody. These cells also bind positively to a cyanobacteria-specific (CYA762) 16*S* ribosomal RNA–targeted oligonucleotide probe (Fig. 2D) which cross-reacts with a large number of cyanobacteria (*16*, *19*). Additionally, 16*S* ribosomal DNA sequencing using cyanobacteria-specific primers (*16*, *20*) yielded a 556-bp sequence (GenBank accession number AY580333) from genomic DNA preparations that match cyanobacterial sequences ($n = 100$) related to *Synechococcus* sp., *Prochlorococcus* sp., or uncultured cyanobacteria in the Order Chroococcales, a paraphyletic group (*21*), with 93 to 97% sequence homology.

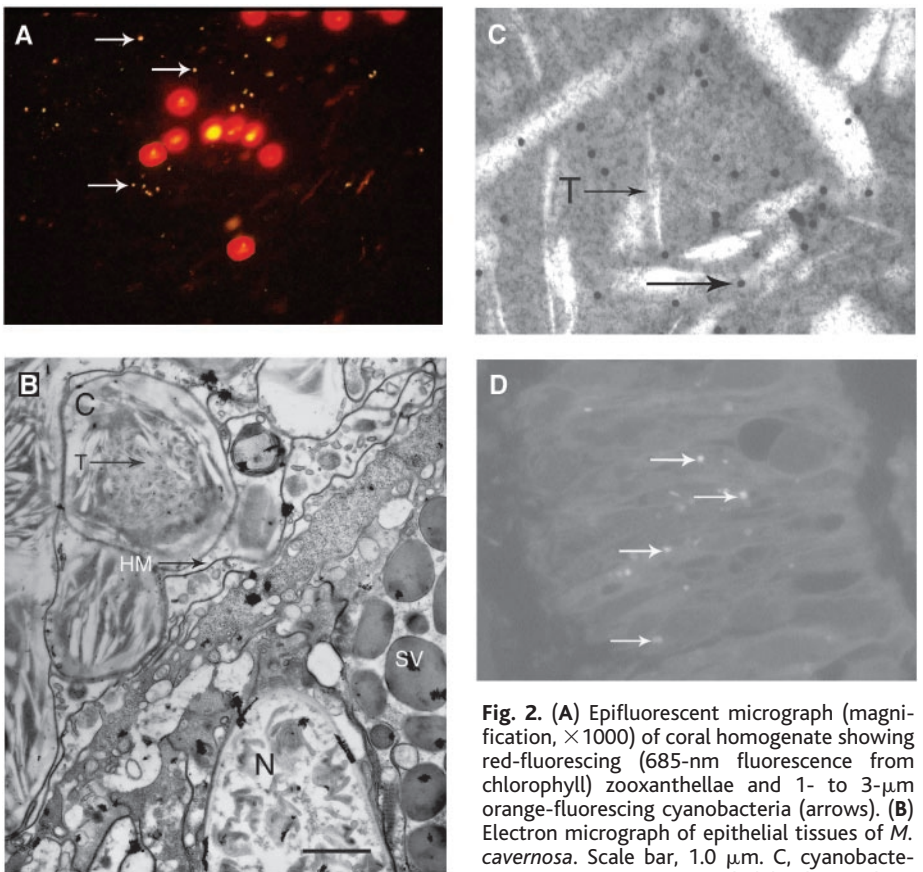


Fig. 2. (**A**) Epifluorescent micrograph (magnification, ×1000) of coral homogenate showing red-fluorescing (685-nm fluorescence from chlorophyll) zooxanthellae and 1- to 3-μm orange-fluorescing cyanobacteria (arrows). (**B**) Electron micrograph of epithelial tissues of *M. cavernosa*. Scale bar, 1.0 μm. C, cyanobacterium; N, nematocyst; T, thylakoid; HM, host membrane; SV, secretory vesicle. (**C**) Immunogold labeling (20-nm gold particles, arrows) of thin sections for phycoerythrin (magnification, ×50,000). (**D**) Fluorescent in situ hybridization micrograph (magnification, ×1000) of *M. cavernosa* epithelial tissues showing positive binding of cyanobacterial-specific probe (arrows).

To assess the potential for nitrogen fixation in this coral, we challenged protein extracts with a polyclonal antibody to the 32-kD Fe protein subunit of nitrogenase (Fig. 3A). The results yielded a single positive cross-reaction, strongly indicating expression of the gene in the coral. Again, using the same antibody to nitrogenase, we used immunogold probing of thin sections (Fig. 3B) and found that the antibody binds significantly (ANOVA, $P = 0.006$) to the cyanobacterial cells [8.8 particles per cell ± 1.3 (SE)] when compared to controls [4.5 particles per cell ± 0.5 (SE)] without the primary antibody. Many members of the Order Chroococcales are capable of fixing nitrogen. Our results clearly suggest that endosymbiotic cyanobacteria capable of fixing nitrogen are present in *M. cavernosa* and form a stable long-term association within host cells. This symbiont could potentially be a source of the limiting element nitrogen for the symbiosis through the release of fixed nitrogen products to the coral host.

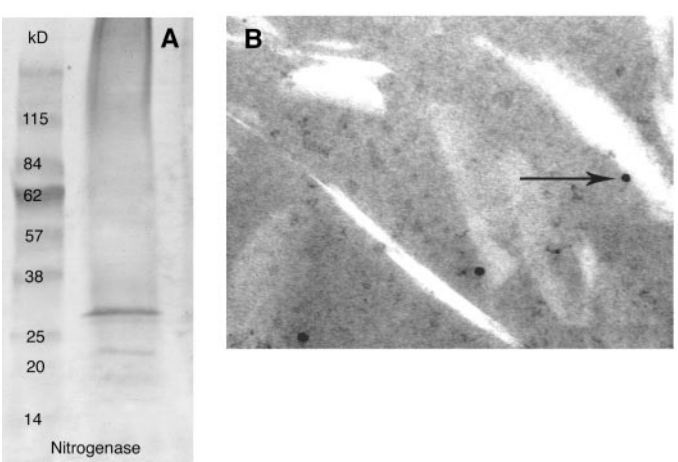


Fig. 3. (**A**) Positive immunoblot for the 32-kD Fe protein of nitrogenase in zooxanthellae-free tissues of *M. cavernosa*. (**B**) Immunogold labeling (20-nm gold particles, arrow) of thin sections for nitrogenase (magnification, ×50,000).

Like the symbiotic cyanobacteria of *M. cavernosa*, free-living cyanobacteria and prochlorophytes that contain phycoerythrin often exhibit strong fluorescence under certain conditions, with a maximum emission between 570 to 580 nm (*22*, *23*). Uncoupled phycoerythrin has been proposed to serve as a storage pool of nitrogen in phycobilin-containing cyanobacteria (*22*) but not in prochlorophytes (*23*). Phycoerythrin detachment from the photosynthetic apparatus in cyanobacteria and prochlorophytes

can be caused by exposure to glycerol and results in strong fluorescence by eliminating the quenching associated with energy transfer from phycoerythrin to the reaction centers (*22*, *23*). The cyanobacterial symbionts of *M. cavernosa* are exposed to high concentrations of glycerol in the coral, because it is the major carbon compound translocated from the symbiotic zooxanthellae to the host tissues (*24*), and this may explain both the unusual ultrastructure and the characteristic orange fluorescence emission of the symbiotic cyanobacteria. Because little or no energy is transferred from phycoerythrin to primary photochemistry, glycerol supplied from the zooxanthellae may serve as an energy source for the cyanobacteria operating heterotrophically and provide a steady supply of reductant and adenosine triphosphate for nitrogen fixation in the symbionts. Nitrogenase is also sensitive to molecular and reactive species of oxygen that accumulate during photosynthesis (*25*, *26*), and the symbiotic cyanobacteria, operating heterotrophically, could quench molecular oxygen via respiration and/or by the Mehler reaction (*27*). Additionally, the coral environment is well suited for the temporal separation of photosynthesis and nitrogen fixation, because coral tissues experience extreme hypoxia at night (*25*). The presence of an additional symbiont in a zooxanthellate coral that is nitrogen-limited (*2*) suggests that nitrogen fixation may be an important supplemental source of the limiting element for the symbiotic association, and it highlights the potential significance of microbial consortia composed of photosynthetic eukaryotes and prokaryotes (*28*). Cyanobacteria are involved in many diverse mutualistic symbioses in both terrestrial and marine environments, and they provide critical ecological services, including important contributions to the global nitrogen cycle (*29*).

References and Notes

1. J. E. N. Veron, *Corals in Space and Time* (Cornell Univ. Press, Ithaca, NY, 1995).
2. P. G. Falkowski, Z. Dubinsky, L. Muscatine, L. R. McCloskey, *Bioscience* **43**, 606 (1993).
3. Z. Dubinsky, P. L. Jokiel, *Pacific Sci.* **48**, 313 (1994).
4. F. Rohwer, M. Breitbart, J. Jara, F. Azam, N. Knowlton, *Coral Reefs* **20**, 85 (2001).
5. F. Rohwer, V. Seguritan, F. Azam, N. Knowlton, *Mar. Ecol. Prog. Ser.* **243**, 1 (2002).
6. H. W. Ducklow, in *Coral Reefs, Ecosystems of the World*, Z. Dubinsky, Ed. (Elsevier, Amsterdam, 1990), pp. 265–290.
7. W. J. Wiebe, R. E. Johannes, K. L. Webb, *Science* **188**, 257 (1975).
8. N. Shashar, Y. Cohen,Y. Loya, N. Star, *Mar. Ecol. Prog. Ser.* **111**, 259 (1994).
9. M. V. Matz *et al.*, *Nature Biotechnol.* **17**, 969 (1999).
10. S. G. Dove, O. Hoegh-Guldberg, S. Ranganathan, *Coral Reefs* **19**, 197 (2001).
11. C. H. Mazel *et al.*, *Limnol. Oceanogr.* **48**, 402 (2003).
12. I. V. Kelmanson, M. V. Matz, *Mol. Biol. Evol.* **20**, 1125 (2003).
13. Y. A. Labas *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 4256 (2002).
14. C. H. Mazel, *Mar. Ecol. Prog. Ser.* **120**, 185 (1995).
15. A. N. Glazer, J. A. West, C. Chan, *Biochem. Syst. Ecol.* **10**, 203 (1982).
16. Materials and methods are available as supporting material on *Science* Online.
17. A. M. Gilmore *et al.*, *Photochem. Photobiol.* **77**, 515 (2003).
18. A. A. Heikal, S. T. Hess, G. S. Baird, R. Y. Tsein, W. W. Webb, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11996 (2000).
19. W. Schönhuber *et al.*, *Appl. Environ. Microbiol.* **65**, 1259 (1999).
20. U. Nübel, F. Garcia-Pichel, G. Muyzer, *Appl. Environ. Microbiol.* **63**, 3327 (1997).
21. M. K. Litvaitis, *Hydrobiologia* **468**, 135 (2002).
22. M. Wyman, R. P. F. Gregory, N. G. Carr, *Science* **230**, 818 (1985).
23. H. Lokstein, C. Steglich, W. R. Hess, *Biochim. Biophys. Acta* **1410**, 97 (1999).
24. L. Muscatine, in *Coral Reefs, Ecosystems of the World*, Z. Dubinsky, Ed. (Elsevier, Amsterdam, 1990), pp. 75–87.
25. M. Kühl, Y. Cohen, T. Dalsgaard, B. B. Jørgensen, N. P. Revsbech, *Mar. Ecol. Prog. Ser.* **117**, 159 (1995).
26. M. P. Lesser, *Coral Reefs* **16**, 187 (1997).
27. I. Berman-Frank *et al.*, *Science* **294**, 1534 (2001).
28. N. Knowlton, F. Rowher, *Am. Nat.* **162**, S51 (2003).
29. D. G. Adams, in *The Ecology of Cyanobacteria*, B. A. Whitton, M. Potts, Eds. (Kluwer, Dordrecht, Netherlands, 2000), pp. 523–561.

30. The authors thank A. Blakeslee, J. H. Farrell, V. A. Kruse, A. Mumford, and E. Sullivan for technical assistance; J. Zehr for the nitrogenase antibody; E. Gantt for the phycoerythrin antibody; and M. Litvaitis for cyanobacterial primers. This work was funded by grants from the Office of Naval Research–Environmental Optics Program, and logistical support was provided by the Caribbean Marine Research Center, Lee Stocking Island, Bahamas. The experiments conducted for this study comply with the current laws of the Bahamas and the United States.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/997/DC1

Materials and Methods

Fig. S1

14 April 2004; accepted 16 July 2004

Modulation of Hematopoietic Stem Cell Homing and Engraftment by CD26

Kent W. Christopherson II,[1,2]* Giao Hangoc,[1,2] Charlie R. Mantel,[1,2] Hal E. Broxmeyer[1,2]†

Hematopoietic stem cell homing and engraftment are crucial to transplantation efficiency, and clinical engraftment is severely compromised when donor-cell numbers are limiting. The peptidase CD26 (DPPIV/dipeptidylpeptidase IV) removes dipeptides from the amino terminus of proteins. We present evidence that endogenous CD26 expression on donor cells negatively regulates homing and engraftment. By inhibition or deletion of CD26, it was possible to increase greatly the efficiency of transplantation. These results suggest that hematopoietic stem cell engraftment is not absolute, as previously suggested, and indicate that improvement of bone marrow transplant efficiency may be possible in the clinic.

The efficiency of hematopoietic stem cell (HSC) transplantation is important when donor-cell numbers are limiting. For example, since the first cord blood transplants (*1–3*), the use of cord blood has been mainly restricted to children, not adults, as a result of apprehension about limited cell numbers. Attempts at ex vivo expansion of stem cells for clinical transplantation have not been encouraging (*4*, *5*). An alternative means to enhance engraftment is to increase HSC homing efficiency to bone marrow (BM) niches. Recently it was suggested that HSCs engrafted mice with absolute efficiency (*6–8*). However, if all HSCs homed with absolute efficiency and engraftment, problems of limiting donor cells would not be a concern for clinical transplantation (*3*). Thus, enhancement of homing and engraftment of HSC is needed if advances in transplantation with limiting numbers of HSC are to be realized. On the basis of our work implicating CD26 in granulocyte colony-stimulating factor (G-CSF)–induced mobilization of HSCs and hematopoietic progenitor cells (HPCs) (*9–11*), we investigated the involvement of CD26 in homing and engraftment. Inhibition or deletion of CD26 on donor cells enhanced short-term homing, long-term engraftment, competitive repopulation, secondary transplantation, and mouse survival, which suggests that CD26 is a novel target for increasing transplantation efficiency.

Mouse bone marrow HSCs were defined as cells within the Sca-1$^+$lin$^-$ population

[1]Department of Microbiology and Immunology and the Walther Oncology Center, Indiana University School of Medicine, Indianapolis, IN 46202, USA. [2]Walther Cancer Institute, Indianapolis, IN 46208, USA.

*Present address: Institute of Molecular Medicine, The University of Texas Health Science Center, Houston, TX 77030, USA.

†To whom correspondence should be addressed: hbroxmey@iupui.edu

(*12*). Using chemotaxis assays, we previously established that Diprotin A (Ile-Pro-Ile) is a specific inhibitor of CD26 (*10*). We show here that Diprotin A–treated C57BL/6 Sca-1^+lin$^-$ BM cells exhibited twofold increases in CXCL12-induced migration (Fig. 1A). CD26-deficient (CD26$^{-/-}$) Sca-1^+lin$^-$ BM cells had up to threefold greater migratory response, compared with control Sca-1^+lin$^-$ BM cells (Fig. 1A). Diprotin A treatment of CD26$^{-/-}$ cells did not further enhance chemotaxis (Fig. 1A). Thus, in vitro migration of Sca-1^+lin$^-$ HSC cells to CXCL12 was enhanced by specific inhibition and even more by the absence of CD26 peptidase activity.

Short-term homing experiments used congenic C57Bl/6 (CD45.2$^+$) and BoyJ (CD45.1$^+$) cells to assess recruitment of transplanted HSCs to BM (*13*). Treatment of 1×10^4 to 2×10^4 sorted Sca-1^+lin$^-$ BM C57Bl/6 donor cells with CD26 inhibitor (Diprotin A) for 15 min before transplant resulted in ninefold increases in homing efficiency in BoyJ recipients compared with untreated cells (Fig. 1B). Transplantation of sorted CD26$^{-/-}$ Sca-1^+lin$^-$ BM cells (*14*) resulted in 11-fold increases in homing efficiency (Fig. 1B). This suggests that inhibition, or loss of CD26 activity, significantly increases homing of sorted Sca-1^+lin$^-$ HSCs in vivo. Pretreatment of 20×10^6 low-density (LD) BM donor cells with CD26 inhibitors resulted in 1.5-fold increases in homing efficiency of C57BL/6 Sca-1^+lin$^-$ cells (within the LDBM donor population) into BoyJ recipient BM 24 hours after transplant (Fig. 1C). Transplantation of CD26$^{-/-}$ cells provided a 2.6-fold increase in homing efficiency (Fig. 1C). Thus, inhibition or loss of CD26 activity in the total LDBM donor unit (containing differentiated cells and progenitors) increases in vivo homing of Sca-1^+lin$^-$ HSCs within the LDBM fraction. The use of LDBM cells more accurately represents clinical protocols than does the use of sorted Sca-1^+lin$^-$ HSCs. Differences in homing efficiency between sorted Sca-1^+lin$^-$ cells and LDBM cells may be partially explained by larger numbers of Sca-1^+lin$^-$ donor cells (3×10^4) contained within the 20×10^6 cell LDBM donor unit or by accessory cells contained within the LDBM, but not sorted, cell population.

Treatment of 10×10^6 LDBM donor cells with CXCR4 antagonist AMD3100 (*15*) for 15 min before transplantation reversed increases in homing efficiency of CD26-inhibited or deleted Sca-1^+lin$^-$ cells (Fig. 1D). AMD3100 treatment itself reduced homing efficiency compared with control cells (Fig. 1D). Migration data from treatment with AMD3100 and in vitro CD26$^{-/-}$ HSC/HPC, combined with our previous studies (*10*), suggest that CXCL12 is a logical downstream target of enhanced transplant efficiency. This is consistent with an important role for CXCL12 in migration (*16*), mobilization (*17–19*), homing, and engraftment of HSCs (*3*, *20–22*), holding HSCs and HPCs in the bone marrow (*23*), and enhancing cell survival, an additional component of HSC engraftment (*24*, *25*).

Although CD26 peptidase activity is rapidly lost after treatment with CD26 inhibitors (Diprotin A or Val-Pyr), recovery begins within 4 hours after treatment (Fig. 1E), which might explain homing and enhancement differences of inhibitor-treated, compared with CD26$^{-/-}$, donor cells (Fig. 1, B and C). As reported for CD34$^+$ cells, cytokine treatment (with interleukin-6 and stem cell factor) of Sca-1^+lin$^-$ cells resulted in increased CXCR4 expression (fig. S1) (*26*). Cytokine treatment did not affect CD26 expression or activity (fig. S1), which suggests that these cytokines do not regulate CD26.

Transplant efficiency requires consideration of long-term donor HSC engraftment. Transplants were performed with 5×10^5 LDBM C57Bl/6 (CD45.2$^+$) or CD26$^{-/-}$ (CD45.2$^+$) cells into lethally irradiated congenic BoyJ (CD45.1$^+$) recipients. CD26$^{-/-}$ donor cells made a significantly greater contribution to peripheral blood (PB) leukocytes 6 months after transplant (Fig. 2A and fig. S2). This was especially apparent at limiting cell dilutions (Fig. 2A) and correlated with changes in mouse survival (Fig. 2, B and C). At day 60, 0% survival was observed with 2.5×10^4 control cells (Fig. 2B); 80% survival was observed with an equivalent number of CD26$^{-/-}$ cells (Fig. 2C). Transplantation of 2.5×10^4 normal cells was below that

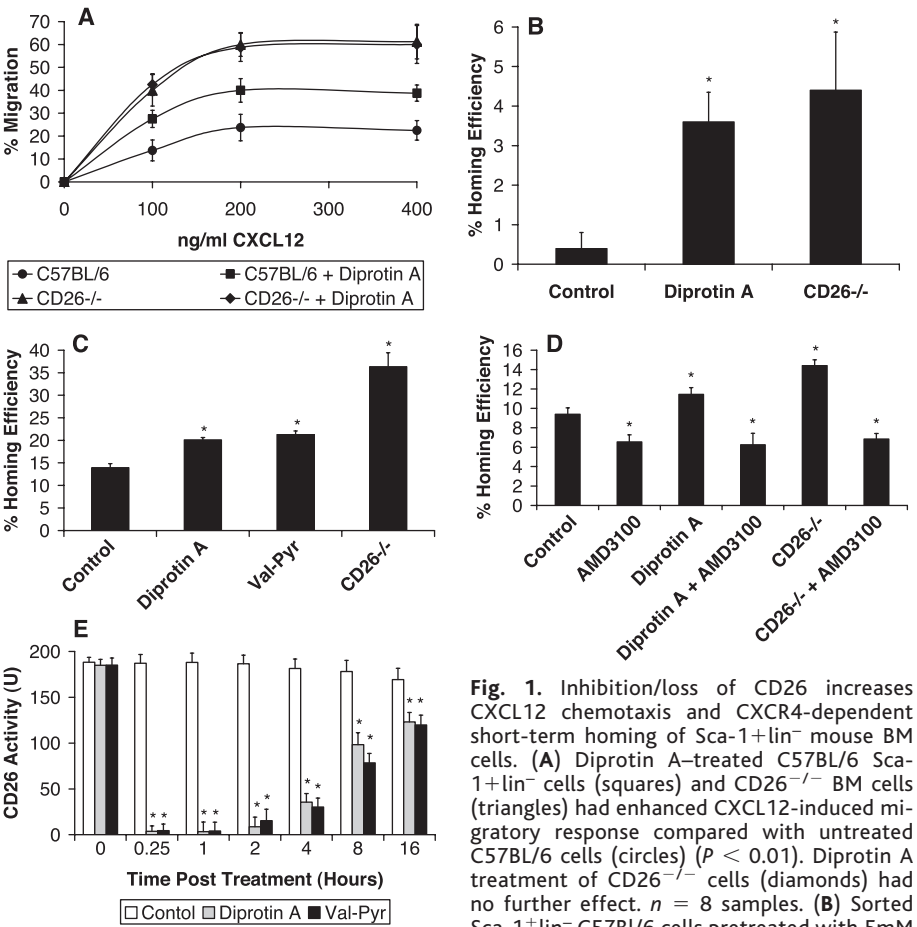


Fig. 1. Inhibition/loss of CD26 increases CXCL12 chemotaxis and CXCR4-dependent short-term homing of Sca-1+lin$^-$ mouse BM cells. (**A**) Diprotin A–treated C57BL/6 Sca-1+lin$^-$ cells (squares) and CD26$^{-/-}$ BM cells (triangles) had enhanced CXCL12-induced migratory response compared with untreated C57BL/6 cells (circles) ($P < 0.01$). Diprotin A treatment of CD26$^{-/-}$ cells (diamonds) had no further effect. $n = 8$ samples. (**B**) Sorted Sca-1^+lin$^-$ C57BL/6 cells pretreated with 5mM Diprotin A for 15 min and CD26$^{-/-}$ cells have increased short-term homing into BoyJ recipient mice ($P < 0.05$; $n = 10$ mice; total from two experiments). (**C**) Sca-1^+lin$^-$ cells within donor LDBM pretreated with CD26 inhibitors (Diprotin A or Val-Pyr) or transplantation of CD26$^{-/-}$ cells significantly increases short-term homing of donor cells ($P < 0.01$; $n = 6$ mice; total from two experiments). (**D**) Increased homing efficiency of Sca-1^+lin$^-$ C57BL/6 HSC within LDBM cells noted with Diprotin A treatment or with CD26$^{-/-}$ cells is reversible by their treatment with CXCR4 antagonist AMD3100 for 15 min before transplant. AMD3100 also reduces homing efficiency of C57BL/6 donor cells in the absence of CD26 inhibition ($P < 0.05$; $n = 5$ mice). (**E**) CD26 peptidase activity (U/1000 cells; 1U = 1 pmol p-nitroanilide per min) of C57BL/6 BM cells is rapidly lost with 15 min inhibitor treatment ($P < 0.01$); recovery begins within 4 hours.

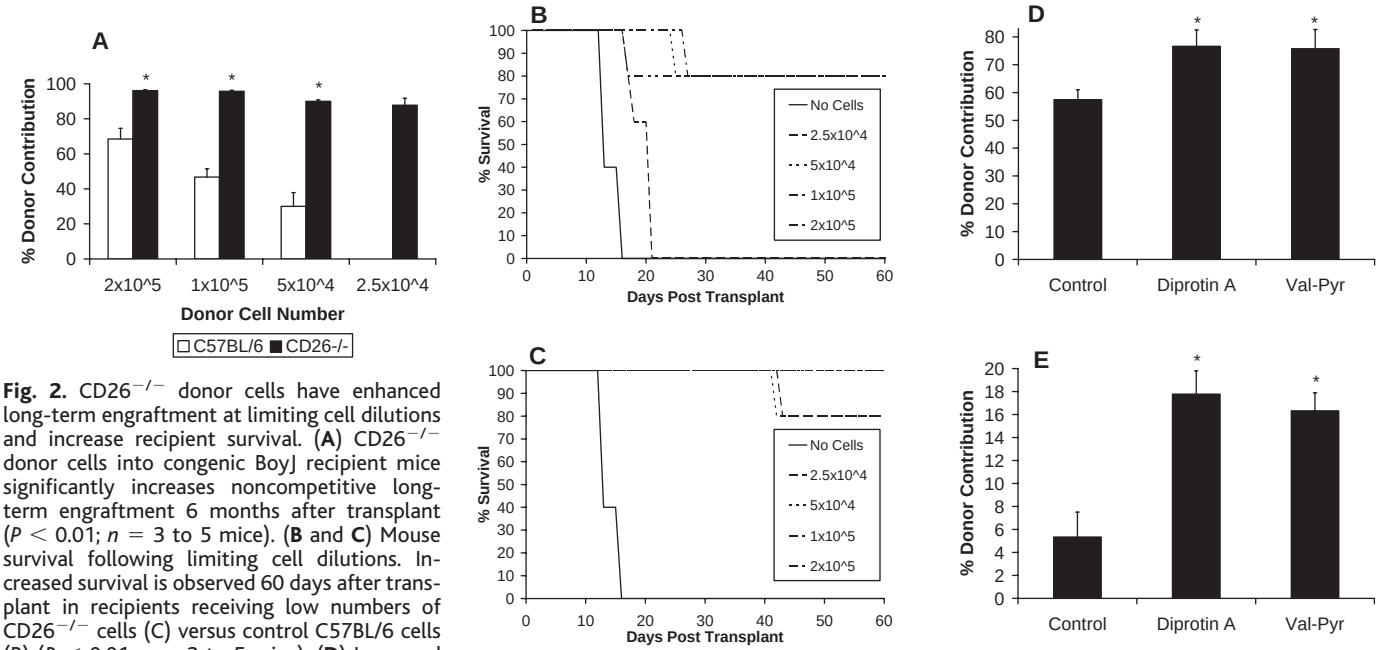


Fig. 2. CD26$^{-/-}$ donor cells have enhanced long-term engraftment at limiting cell dilutions and increase recipient survival. (**A**) CD26$^{-/-}$ donor cells into congenic BoyJ recipient mice significantly increases noncompetitive long-term engraftment 6 months after transplant ($P < 0.01$; $n = 3$ to 5 mice). (**B** and **C**) Mouse survival following limiting cell dilutions. Increased survival is observed 60 days after transplant in recipients receiving low numbers of CD26$^{-/-}$ cells (C) versus control C57BL/6 cells (B) ($P < 0.01$; $n = 3$ to 5 mice). (**D**) Increased donor-cell contribution to formation of PB leukocytes during noncompetitive long-term engraftment assays was also observed with CD26 inhibitor treatment (Diprotin A or Val-Pyr) 6 months after transplant ($P < 0.05$; $n = 5$ mice). (**E**) Even greater donor contribution to chimerism was observed in 6-month secondary transplants receiving cells from primary mice engrafted with CD26-inhibited cells ($P < 0.01$; $n = 5$ mice).

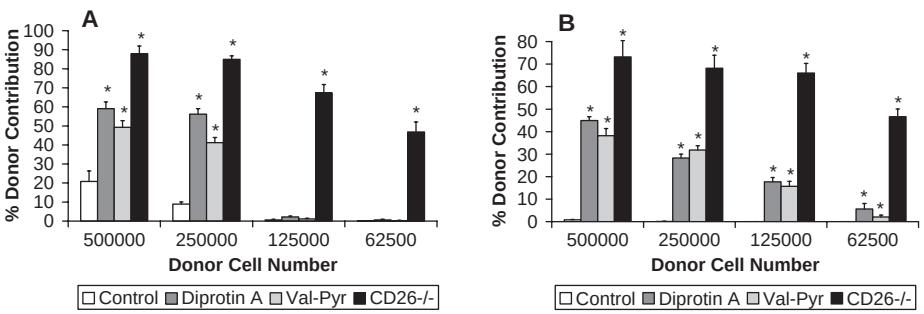


Fig. 3. CD26-inhibited and CD26$^{-/-}$ cells have increased competitive repopulation and engraftment of secondary repopulating HSCs compared with control cells. (**A**) Increased donor-cell chimerism in competitive repopulation assays was observed with Diprotin A or Val-Pyr treatment of 5×10^5 or 2.5×10^5 donor cells in direct competition with 5×10^5 competitor recipient BoyJ cells ($P < 0.01$). CD26$^{-/-}$ cells had a greatly enhanced contribution to chimerism at all donor-cell numbers ($P < 0.01$; $n = 5$ mice). (**B**) An even greater increase in donor-cell contribution was observed with CD26 inhibitor–treated donors and CD26$^{-/-}$ donors in secondary transplanted recipient BoyJ mice ($P \leq 0.01$; $n = 5$).

required for mouse survival. Recipient survival is dependent on short- and long-term reconstitution of marrow. The absence of surviving mice in this group by day 21 suggests that loss of short-term reconstitution may be responsible for lethality at this donor-cell dose. At limiting transplanted donor cells, long-term engraftment and mouse survival increased with CD26$^{-/-}$ donor cells. At nonlimiting donor-cell numbers (2×10^5), improvement is also observed in engraftment and survival with CD26$^{-/-}$ cells at day 60, which suggests that long-term reconstitution is also targeted.

At nonlimiting cell doses and in a noncompetitive assay, treatment of C57BL/6 donor cells with either CD26 inhibitor resulted in a one-third increase in donor-cell contribution to leukocyte formation in lethally irradiated BoyJ recipients relative to untreated cells (Fig. 2D). In secondary transplanted recipient mice, a threefold increase in donor-cell contribution to PB leukocytes was seen with CD26 inhibition (Fig. 2E and fig. S3). Increases in secondary repopulating HSCs compared with repopulating HSCs in primary recipients indicates an increased homing/engraftment of self-renewing stem cells with CD26 inhibition.

Competitive repopulating HSC assays provide the most functional assessment of HSC by direct comparison of engraftment from experimental donor cells (CD45.2$^+$) relative to constant numbers of competitor cells (CD45.1$^+$) (*13*, *27*). Six months after transplant, increased donor contribution to chimerism was observed with Diprotin A or Val-Pyr treatment relative to cotransplanted cells (Fig. 3A and fig. S4A). At limiting donor-cell numbers (1.25×10^5 and 0.625×10^5), no significant increases in donor contribution were observed with CD26 inhibitor treatment (Fig. 3A). However, CD26$^{-/-}$ donor cells significantly enhanced chimerism at all donor-cell numbers measured (Fig. 3A). Even greater increases in donor-cell contribution were observed with CD26 inhibitor–treated and CD26$^{-/-}$ donor cells in secondary transplanted BoyJ recipients 4 months after transplant (Fig. 3B); this result was more striking when CD26$^{-/-}$ donor cells were used (Fig. 3B and fig. S4B). It is unlikely that some increases in CD26$^{-/-}$ donor-cell engraftment are the result of increased HSC cell numbers in this population. Numbers of Sca-1$^+$lin$^-$ cells ($2.69 \pm 0.49 \times 10^4$ per femur pair in C57BL/6 and $2.75 \pm 0.15 \times 10^4$ per femur pair in CD26$^{-/-}$) and CFU-GM, BFU-E, and CFU-GEMM in PB, BM, and spleen (*11*) are comparable in CD26$^{-/-}$ and control C57BL/6 mice. Cycling status of CD26$^{-/-}$ cells was examined because HSCs/HPCs not in G_0/G_1 are reported to manifest decreased homing and engraftment (*28*). No significant differences were seen in cycling status between CD26$^{-/-}$ and control C57BL/6 Sca-1$^+$lin$^-$ BM cells, either when freshly isolated or after 24 hours of preincubation with growth factors (fig. S5).

Enhancing transplant efficiency has clinical implication but is being debated. Recent reports suggest that HSCs engraft mice with absolute efficiency (*7*, *8*). One report (*7*) was heavily influenced by mathematical correction factors, and the other (*8*) addressed single-cell transplants by a subset of HSCs among competitor cells that themselves could save the lethally irradiated recipient. Contrary to this is the reality of the clinical situation (*3*) and studies in which injection of HSCs directly into BM showed enhanced engraftment compared with intravenous administering of cells (*29–31*). Removal of endogenous CD26 activity on donor HSCs increased homing and engraftment. Thus, improvement in transplant efficiency is possible. Further advancement may require more effective use of CD26 inhibitors, which may translate into the use of HSCs for clinical transplantation from sources containing limiting cell numbers, such as cord blood.

References and Notes

1. H. E. Broxmeyer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 3828 (1989).
2. E. Gluckman *et al.*, *N. Engl. J. Med.* **321**, 1174 (1989).
3. H. E. Broxmeyer, F. Smith, in *Thomas' Hematopoietic Cell Transplantation*, K. G. Blume, S. J. Forman, F. R. Appelbaum, Eds. (Blackwell Science, Oxford; Malden, MA, 2004), chap. 43, pp. 550–564.
4. E. J. Shpall *et al.*, *Biol. Blood Marrow Transplant.* **8**, 368 (2002).
5. J. Jaroscak *et al.*, *Blood* **101**, 5061 (2003).
6. H. Ema, H. Nakauchi, *Immunity* **20**, 1 (2004).
7. P. Benveniste, C. Cantin, D. Hyam, N. N. Iscove, *Nature Immunol.* **4**, 708 (2003).
8. Y. Matsuzaki, K. Kinjo, R. C. Mulligan, H. Okano, *Immunity* **20**, 87 (2004).
9. K. W. Christopherson 2nd, G. Hangoc, H. E. Broxmeyer, *J. Immunol.* **169**, 7000 (2002).
10. K. W. Christopherson 2nd, S. Cooper, H. E. Broxmeyer, *Blood* **101**, 4680 (2003).
11. K. W. Christopherson 2nd, S. Cooper, G. Hangoc, H. E. Broxmeyer, *Exp. Hematol.* **31**, 1126 (2003).
12. G. J. Spangrude, S. Heimfeld, I. L. Weissman, *Science* **241**, 58 (1988).
13. D. E. Harrison, *Blood* **55**, 77 (1980).
14. Materials and methods are available as supporting material on *Science* Online.
15. M. M. Rosenkilde *et al.*, *J. Biol. Chem.* **279**, 3033 (2004).
16. D. E. Wright, E. P. Bowman, A. J. Wagers, E. C. Butcher, I. L. Weissman, *J. Exp. Med.* **195**, 1145 (2002).
17. K. Hattori *et al.*, *Blood* **97**, 3354 (2001).
18. H. E. Broxmeyer *et al.*, *Blood* **100**, 609a (abstr. no. 2397) (2002).
19. C. W. Liles *et al.*, *Blood* **102**, 2728 (2003).
20. A. Peled *et al.*, *Science* **283**, 845 (1999).
21. H. E. Broxmeyer, *Int. J. Hematol.* **74**, 9 (2001).
22. T. Ara *et al.*, *Immunity* **19**, 257 (2003).
23. C. H. Kim, H. E. Broxmeyer, *Blood* **91**, 100 (1998).
24. H. E. Broxmeyer *et al.*, *J. Immunol.* **170**, 421 (2003).
25. H. E. Broxmeyer *et al.*, *J. Leukoc. Biol.* **73**, 630 (2003).
26. O. Kollet *et al.*, *Blood* **100**, 2778 (2002).
27. D. E. Harrison, C. T. Jordan, R. K. Zhong, C. M. Astle, *Exp. Hematol.* **21**, 206 (1993).
28. A. Gothot, J. C. van der Loo, D. W. Clapp, E. F. Srour, *Blood* **92**, 2641 (1998).
29. T. Yahata *et al.*, *Blood* **101**, 2905 (2003).
30. J. Wang *et al.*, *Blood* **101**, 2924 (2003).
31. F. Mazurier, M. Doedens, O. I. Gan, J. E. Dick, *Nature Med.* **9**, 959 (2003).

32. These studies were supported by U.S. Public Health Science Grants R01 DK53674, R01 HL67384, and R01 HL56416 to H.E.B. K.W.C. was supported sequentially during these studies by NIH training grant T32 DK07519 to H.E.B. and by a Fellow Award from the Leukemia and Lymphoma Society to K.W.C.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/1000/DC1
Materials and Methods
Figs. S1 to S5
References

23 February 2004; accepted 15 July 2004

Natural Antibiotic Function of a Human Gastric Mucin Against *Helicobacter pylori* Infection

Masatomo Kawakubo,[1,2] Yuki Ito,[1] Yukie Okimura,[2] Motohiro Kobayashi,[1,4] Kyoko Sakura,[2] Susumu Kasama,[1] Michiko N. Fukuda,[4] Minoru Fukuda,[4] Tsutomu Katsuyama,[2] Jun Nakayama[1,3]*

Helicobacter pylori infects the stomachs of nearly a half the human population, yet most infected individuals remain asymptomatic, which suggests that there is a host defense against this bacterium. Because *H. pylori* is rarely found in deeper portions of the gastric mucosa, where *O*-glycans are expressed that have terminal α1,4-linked *N*-acetylglucosamine, we tested whether these *O*-glycans might affect *H. pylori* growth. Here, we report that these *O*-glycans have antimicrobial activity against *H. pylori*, inhibiting its biosynthesis of cholesteryl-α-D-glucopyranoside, a major cell wall component. Thus, the unique *O*-glycans in gastric mucin appeared to function as a natural antibiotic, protecting the host from *H. pylori* infection.

Helicobacter pylori colonizes the gastric mucosa of about half the world's population and is considered a leading cause of gastric malignancies (*1–3*). However, most infected individuals remain asymptomatic or are affected merely by chronic active gastritis (*2*). Only a fraction of infected patients develop peptic ulcer, gastric cancer, and malignant lymphoma. This suggests the presence of host defense mechanisms against *H. pylori* pathogenesis.

Gastric mucins are classified into two types based on their histochemical properties (*4*). The first is a surface mucous cell–type mucin, secreted from the surface mucous cells. The second is found in deeper portions of the mucosa and is secreted by gland mucous cells, including mucous neck cells, cardiac gland cells, and pyloric gland cells.

In *H. pylori* infection, the bacteria are associated solely with surface mucous cell–type mucin (*5*), and two carbohydrate structures, Lewis b and sialyl dimeric Lewis X in surface mucous cells, serve as specific ligands for *H. pylori* adhesins, BabA and SabA, respectively (*6, 7*). *H. pylori* rarely colonizes the deeper portions of gastric mucosa, where the gland mucous cells produce mucins having terminal α1,4-linked *N*-acetylglucosamine (α1,4-GlcNAc) residues attached to core 2–branched *O*-glycans [GlcNAcα1→4Galβ1→4GlcNAcβ1→6 (GlcNAcα1→ 4Galβ1→3)GalNAcα→Ser/Thr] (*8*). Development of pyloric gland atrophy enhances the risk of peptic ulcer or gastric cancer two- to three-fold compared with chronic gastritis without pyloric gland atrophy (*3*). These findings raise the possibility that α1,4-GlcNAc–capped *O*-glycans have protective properties against *H. pylori* infection.

To test this hypothesis, we generated mucin-type glycoproteins containing terminal α1,4-GlcNAc and determined its effect on *H. pylori* in vitro. Because CD43 serves as a preferential core protein of these *O*-glycans (*8*), we generated recombinant soluble CD43 having α1,4-GlcNAc–capped *O*-glycans in transfected Chinese hamster ovary cells (*9*). Soluble CD43 without α1,4-GlcNAc was used as a control.

H. pylori (ATCC43504), incubated with the medium containing varying amounts of recombinant soluble CD43, showed little growth during the first 2.5 days, irrespective of the presence or absence of α1,4-

[1]Department of Pathology and [2]Department of Laboratory Medicine, Shinshu University School of Medicine, and [3]Institute of Organ Transplants, Reconstructive Medicine and Tissue Engineering, Shinshu University Graduate School of Medicine, Asahi 3-1-1, Matsumoto 390-8621, Japan. [4]Glycobiology Program, Cancer Research Center, The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA.

*To whom correspondence should be addressed. E-mail: jun@hsp.md.shinshu-u.ac.jp

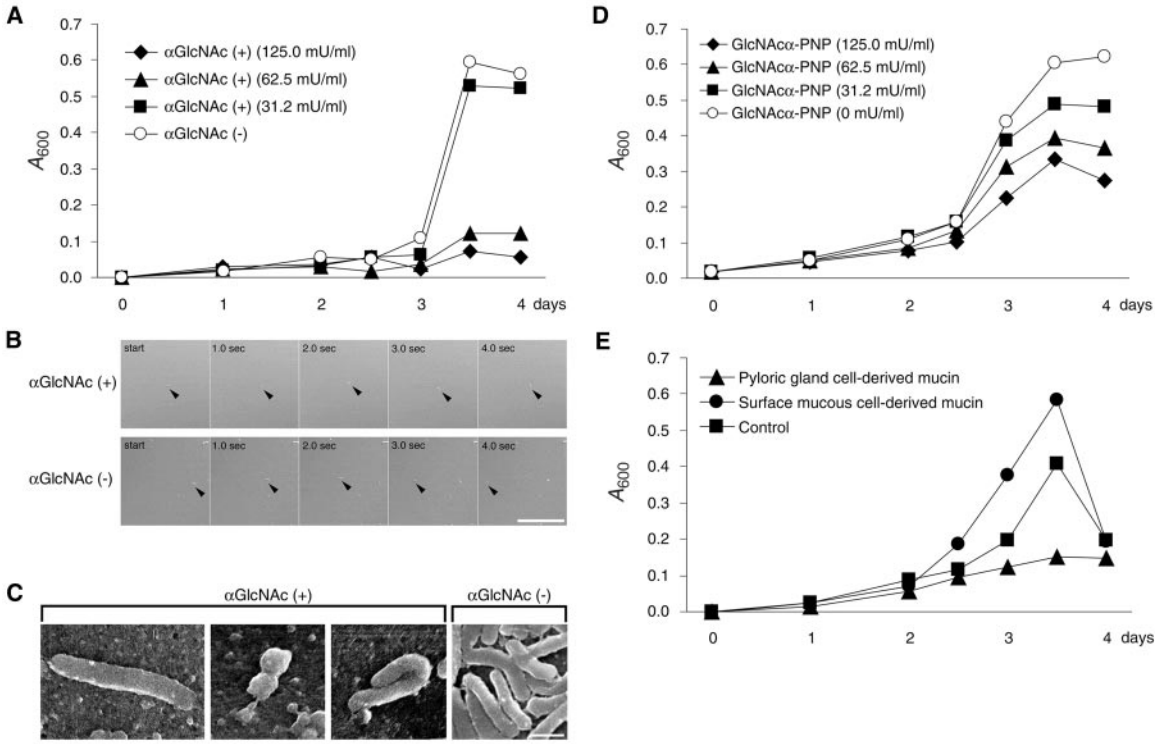


Fig. 1. α1,4-GlcNAc–capped *O*-glycans inhibit the growth and motility of *H. pylori*. (**A**) Growth curves of *H. pylori* cultured in the presence of soluble CD43 with terminal α1,4-GlcNAc [αGlcNAc (+)] or soluble CD43 without terminal α1,4-GlcNAc [αGlcNAc (–)]; the protein concentration of αGlcNAc (–) was the same as that of 125.0 mU/ml of αGlcNAc (+). One miliunit of αGlcNAc (+) corresponds to 1 μg (2.9 nmol) of GlcNAcα-PNP. A_{600}, absorbance at 600 nm. (**B**) Motility of *H. pylori* cultured with 31.2 mU/ml of αGlcNAc (+) or the same protein concentration of αGlcNAc (–) for 3 days by time-lapse recording with 1-s intervals. Representative *H. pylori* is indicated by arrowheads. The mean velocity of seven *H. pylori* cultured in the presence of αGlcNAc (+) and αGlcNAc (–) is 3.1 ± 3.5 μm/s (mean ± SD) and 21.2 ± 2.6 μm/s ($P < 0.001$). Scale bar, 50 μm. (**C**) Scanning electron micrographs of *H. pylori* incubated with 31.2 mU/ml of αGlcNAc (+) or the same protein concentration of αGlcNAc (–) for 3 days. Note abnormal morphologies such as elongation, segmental narrowing, and folding in the culture with αGlcNAc (+). All photographs were taken at the same magnification. Scale bar, 1 μm. (**D**) Growth curves of *H. pylori* cultured in the medium supplemented with various amounts of GlcNAcα-PNP. Growth of the bacteria is suppressed by GlcNAcα-PNP in a dose-dependent manner. (**E**) Growth curves of *H. pylori* cultured in the medium supplemented with pyloric gland cell-derived mucin containing 125 mU/ml of α1,4-GlcNAc or the same protein concentration of surface mucous cell-derived mucin isolated from the human gastric mucosa. The death phase started from 3.5 days, and saline instead of each mucin was supplemented as a control experiment. In (A), (D), and (E), each value represents the average of duplicate measurements.

GlcNAc–capped *O*-glycans, characteristic of the lag phase of *H. pylori* growth (Fig. 1A). After 3 days, microbes cultured in the presence of control soluble CD43 grew rapidly, corresponding to the log phase of bacterial growth. In contrast, soluble CD43 containing more than 62.5 mU/ml of terminal α1,4-GlcNAc impaired log-phase growth. Although growth inhibition was not obvious at a lower concentration (31.2 mU/ml), time-lapse images of the microbes revealed significant reduction of motility under this condition (Fig. 1B). Morphologic examination at the lower concentration revealed abnormalities of the microbe, such as elongation, segmental narrowing, and folding (Fig. 1C). These morphologic changes are distinct from conversion to coccoid form, because reduction of growth, associated with conversion from the bacillary to the coccoid form (*10*), was not apparent under these conditions. These inhibitory effects of soluble CD43 containing terminal α1,4-GlcNAc were also detected against various *H. pylori* strains, including another authentic strain, ATCC43526, and three clinical isolates with a minimum inhibitory concentration between 15.6 mU/ml and 125.0 mU/ml. By contrast, neither inhibitory growth nor abnormal morphology of *H. pylori* was observed at any concentrations of soluble CD43 lacking α1,4-GlcNAc (Fig. 1, A to C). These results indicate that α1,4-GlcNAc–capped *O*-glycans specifically suppress the growth of *H. pylori* in a manner similar to other antimicrobial agents. Similar inhibitory effects on *H. pylori* were also found in another mucin-like glycoprotein, CD34 (*11*) having terminal α1,4-GlcNAc (*12*). In addition, *p*-nitrophenyl-α-*N*-acetylglucosamine (GlcNAcα-PNP) suppressed the growth of *H. pylori* in a dose-dependent manner (Fig. 1D), although the effects were not as strong with soluble CD43 having terminal α1,4-GlcNAc (Fig. 1A). These results provide evidence that the terminal α1,4-GlcNAc residues, rather than scaffold proteins, are critical for growth inhibitory activity against *H. pylori*, and that the presentation of multiple terminal α1,4-GlcNAc residues as a cluster on mucin-type glycoprotein may be important for achieving the optimal activity.

To determine whether natural gastric mucins containing terminal α1,4-GlcNAc can also inhibit growth of *H. pylori*, subsets of human gastric mucins were prepared from the surface mucous cells and pyloric gland cells (*9*). The growth of *H. pylori* was significantly suppressed with mucin derived from pyloric gland cells at 125.0 mU/ml during the log phase (Fig. 1E). A similar inhibitory effect was also observed when the glandular mucin prepared from human gastric juice was tested (*13*). By contrast, mucin derived from surface mucous cells, MUC5AC, stimulated growth. These results support the hypothesis that natural gastric mucins containing terminal α1,4-GlcNAc, secreted from gland mucous cells, have antimicrobial activity against *H. pylori*.

The morphologic abnormalities of *H. pylori* induced by α1,4-GlcNAc–capped *O*-glycans are similar to those induced by antibiotics such as β-lactamase inhibitors, which disrupt biosynthesis of peptidoglycan in the cell wall (*14, 15*). Therefore, these *O*-glycans may inhibit cell wall biosynthesis in *H. pylori*. The cell wall of

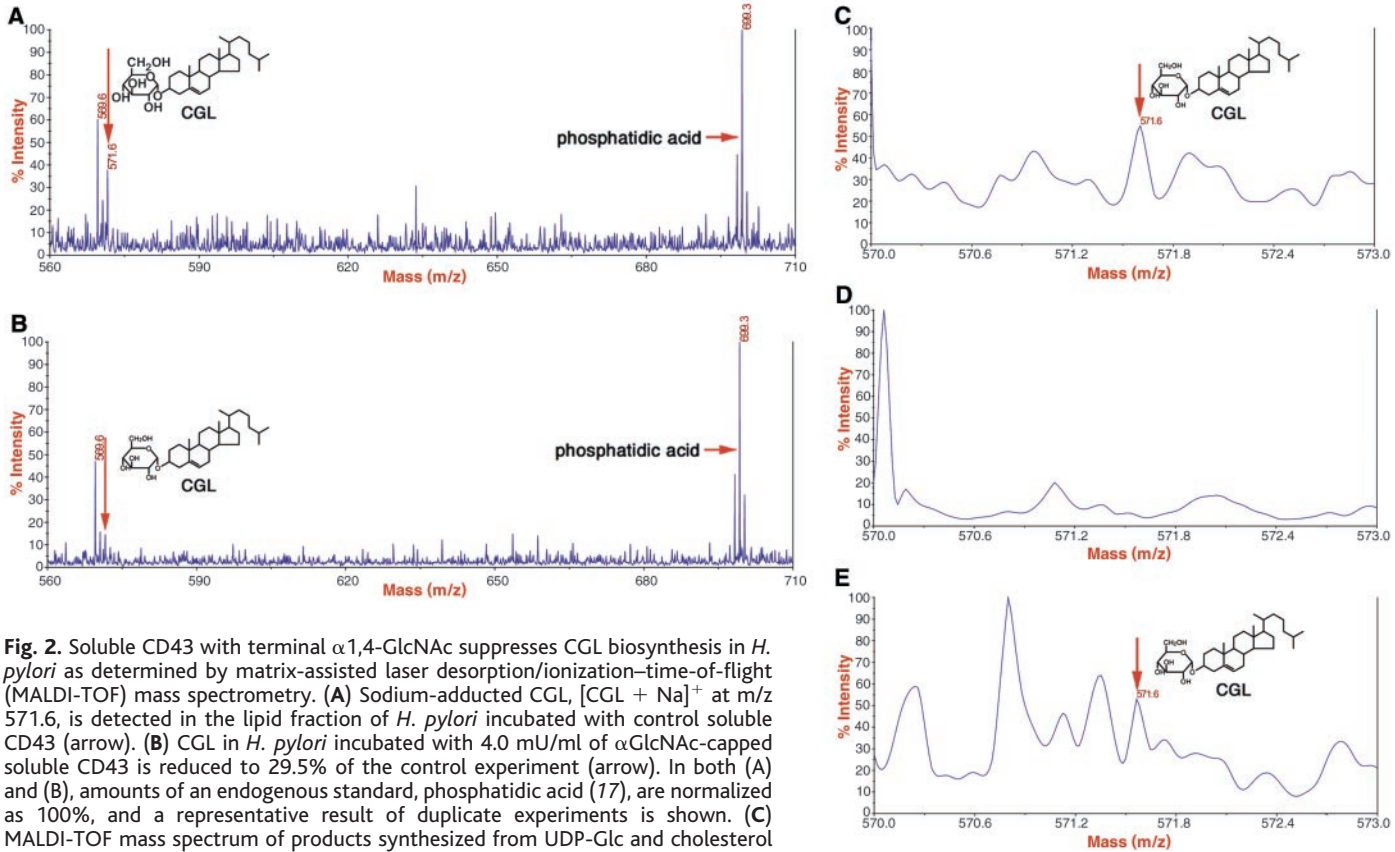


Fig. 2. Soluble CD43 with terminal α1,4-GlcNAc suppresses CGL biosynthesis in *H. pylori* as determined by matrix-assisted laser desorption/ionization–time-of-flight (MALDI-TOF) mass spectrometry. (**A**) Sodium-adducted CGL, $[CGL + Na]^+$ at m/z 571.6, is detected in the lipid fraction of *H. pylori* incubated with control soluble CD43 (arrow). (**B**) CGL in *H. pylori* incubated with 4.0 mU/ml of αGlcNAc-capped soluble CD43 is reduced to 29.5% of the control experiment (arrow). In both (A) and (B), amounts of an endogenous standard, phosphatidic acid (*17*), are normalized as 100%, and a representative result of duplicate experiments is shown. (**C**) MALDI-TOF mass spectrum of products synthesized from UDP-Glc and cholesterol by sonicated *H. pylori*. $[CGL + Na]^+$ at m/z 571.6 is shown. (**D** and **E**) Mass spectrum of products synthesized from UDP-Glc and cholesterol by sonicated *H. pylori* in the presence of 50.0 mU/ml of α1,4-GlcNAc–capped soluble CD43 (D) or control soluble CD43 (E). Note that CGL is not synthesized in the presence of α1,4-GlcNAc–capped soluble CD43 in (D).

Fig. 3. Absence of α-CGs including CAG, CGL, and CPG in *H. pylori* cultured without exogenous cholesterol. Total glycolipids extracted from *H. pylori* incubated with Brucella broth lacking cholesterol (lane 1) or containing 0.005% cholesterol (lane 2) were analyzed by thin-layer chromatography.

Helicobacter species characteristically contains α-cholesteryl glucosides (α-CGs), of which the major components are cholesteryl-α-D-glucopyranoside (CGL), cholesteryl-6-*O*-tetradecanoyl-α-D-glucopyranoside (CAG), and cholesteryl-6-*O*-phosphatidyl-α-D-glucopyranoside (CPG) (*16*). Mass spectrometric analysis of the cell wall components from *H. pylori* cultured with α1,4-GlcNAc–capped *O*-glycans displayed reduced lipid-extractable cell wall constituents (Fig. 2B). In particular, the levels of CGL, relative to phosphatidic acid (*17*), were significantly reduced as compared with controls (Fig. 2, A and B). These results suggest that α1,4-GlcNAc–capped *O*-glycans directly inhibit biosynthesis of CGL in vivo by *H. pylori*.

CGL is likely formed by a UDP-Glc:sterol α-glucosyltransferase, which transfers glucose (Glc) from UDP-Glc to the C3 position of cholesterol with α-linkage. Incubation of cholesterol and UDP-Glc with *H. pylori* lysates revealed substantial amounts of CGL by mass spectrometry (Fig. 2C), demonstrating the activity of UDP-Glc:sterol α-glucosyltransferase in *H. pylori*. When soluble CD43 containing terminal α1,4-GlcNAc was added to this assay, production of CGL was suppressed (Fig. 2D), whereas no effect was seen with control soluble CD43 (Fig. 2E). Considering structural similarity between α-linked GlcNAc found in the gland mucous cell–type mucin and the α-linked Glc found in CGL, these findings suggest that the terminal α1,4-GlcNAc residues could directly inhibit the α-glucosyltransferase activity through an end-product inhibition mechanism (*18*), resulting in decreased CGL biosynthesis.

Genes involved in the biosynthesis of cholesterol are not found in the genome database of *H. pylori* (*19*). Thus, *H. pylori* may not be able to synthesize CGL in the absence of exogenous cholesterol. When *H. pylori* was cultured for 5 days without cholesterol, bacterial growth was significantly reduced (table S1). In such cultures, *H. pylori* was elongated and no motile microbes were found. When *H. pylori* was further cultured without cholesterol for up to 21 days, the microbes died off completely. By contrast, when *H. pylori* was cultured with cholesterol, bacteria grew well, and no signs of abnormality were detected (table S1). *H. pylori* cultured with cholesterol (*9*) revealed a typical triplet of α-CGs including CGL (Fig. 3, lane 2), while α-CGs were not detected in *H. pylori* cultured without cholesterol (Fig. 3, lane 1). Moreover, no antibacterial effect of soluble CD43 containing terminal α1,4-GlcNAc was observed on bacterial strains lacking CGL such as *Escherichia coli*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, *Staphylococcus aureus*, α-*Streptococcus*, and *Streptococcus pneumoniae* (*9*). These results collectively indicate that synthesis of CGL by using exogenously supplied cholesterol is required for the survival of *H. pylori* and that antimicrobial activity of α1,4-GlcNAc–capped *O*-glycans may be restricted to bacterial strains expressing CGL.

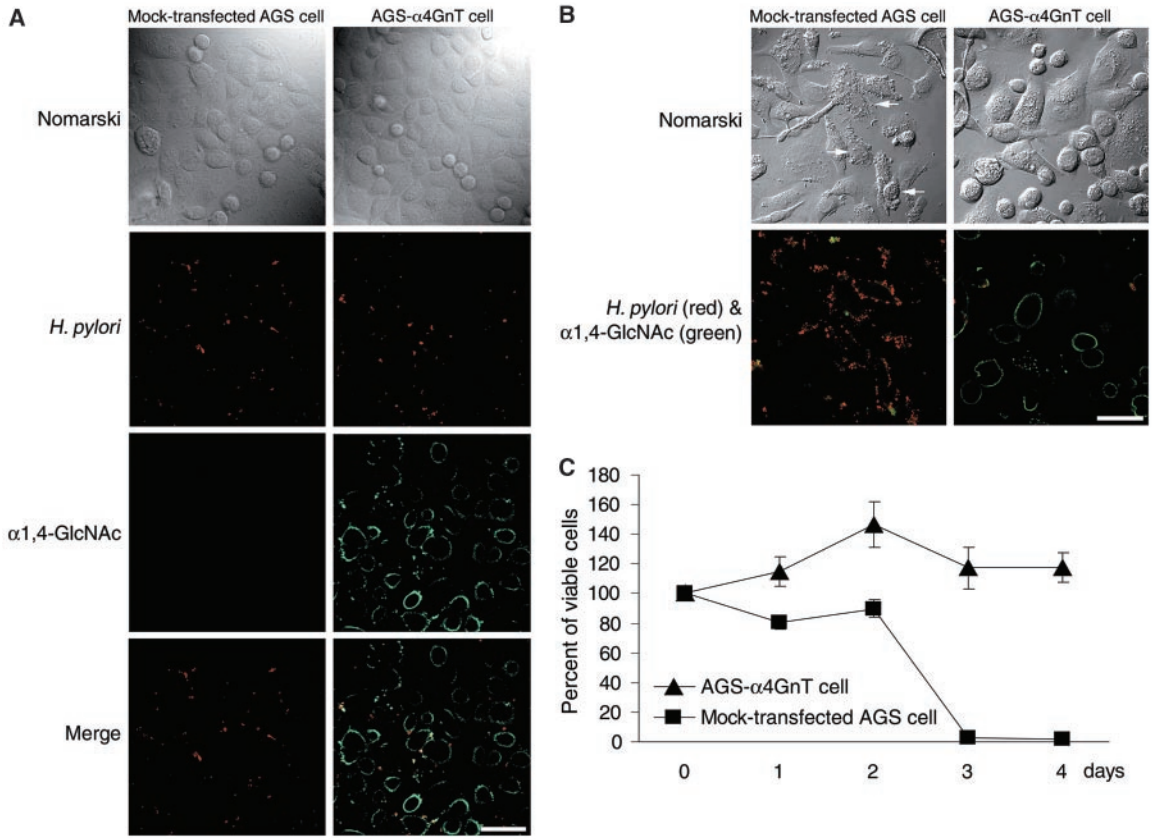


Fig. 4. α1,4-GlcNAc–capped *O*-glycans protect the host cells. AGS cells were incubated with *H. pylori* for 8 hours (**A**) or 24 hours (**B**), and doubly stained with anti-*H. pylori* antibody (red) and HIK1083 antibody specific for terminal α1,4-GlcNAc (*27*) (green). (A) Note that comparable number of *H. pylori* adhered to both mock-transfected AGS cells and AGS-α4GnT cells. (B) After 24 hours, marked damage such as cell flatness or shrinkage are noted (arrows) in mock-transfected AGS cells; no cellular damage and few attached bacteria are found in AGS-α4GnT cells. (Top) Nomarski photographs of the same field. Scale bar, 50 μm. (**C**) Viabilities of AGS cells cocultured with *H. pylori* for 4 days determined by MTS assay. Note that viability of mock-transfected AGS cells was significantly reduced after the third day, whereas AGS-α4GnT cells were fully viable for up to 4 days. The assay was done with triplicate measurements, and error bars indicate SD.

To test whether mucous cells expressing α1,4-GlcNAc–capped *O*-glycans protect themselves against *H. pylori* infection, gastric adenocarcinoma AGS-α4GnT cells stably transfected with α4GnT cDNA were cocultured with *H. pylori* (*9*). With a short-term incubation (8 hours), the microbes attached equally well to AGS-α4GnT cells and mock-transfected AGS cells. No significant damage was observed in either group of cells (Fig. 4A). Upon prolonged incubation (24 hours), mock-transfected AGS cells exhibited remarkable deterioration, such as flatness or shrinkage, with increased number of associated *H. pylori* (Fig. 4B), and the number of viable AGS cells was dramatically reduced after the third day (Fig. 4C). This cellular damage may be attributed to the perturbed signal transduction in AGS cells, where a tyrosin phosphatase, SHP-2, is constitutively activated by *H. pylori* CagA protein (*20*). By contrast, growth of *H. pylori* in cultures with AGS-α4GnT cells was markedly suppressed, and cellular damage found in mock-transfected AGS cells was barely detected in these cells (Fig. 4B). Thus, the viability of AGS-α4GnT cells was fully maintained for up to 4 days (Fig. 4C). These results indicate that α1,4-GlcNAc–capped *O*-glycans have no effect on the adhesion of *H. pylori* to AGS-α4GnT cells, but protect the host cells from *H. pylori* infection.

Glycan chains play diverse roles as ligands for cell surface receptors (*11, 21–23*) and as modulators of receptors and adhesive proteins (*24–26*). The present study reveals a new aspect of mammalian glycan function as a natural antibiotic. Because α1,4-GlcNAc–capped *O*-glycans are produced by human gastric gland mucous cells, the present study provides a basis for development of novel and potentially safe therapeutic agents to prevent and treat *H. pylori* infection in humans without adverse reactions.

References and Notes

1. R. M. Peek Jr., M. J. Blaser, *Nature Rev. Cancer* **2**, 28 (2002).
2. D. R. Cave, *Semin. Gastrointest. Dis.* **12**, 196 (2001).
3. P. Sipponen, H. Hyvarinen, *Scand. J. Gastroenterol. Suppl.* **196**, 3 (1993).
4. H. Ota *et al.*, *Histochem. J.* **23**, 22 (1991).
5. E. Hidaka *et al.*, *Gut* **49**, 474 (2001).
6. D. Ilver *et al.*, *Science* **279**, 373 (1998).
7. J. Mahdavi *et al.*, *Science* **297**, 573 (2002).
8. J. Nakayama *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8991 (1999).
9. Materials and methods are available as supplemental material on *Science* Online.
10. H. Enroth *et al.*, *Helicobacter* **4**, 7 (1999).
11. J.-C. Yeh *et al.*, *Cell* **105**, 957 (2001).
12. M. Kawakubo, J. Nakayama, unpublished observations.
13. Y. Ito, M. Kawakubo, J. Nakayama, unpublished observations.
14. T. Horii *et al.*, *Helicobacter* **7**, 39 (2002).
15. J. Finlay, L. Miller, J. A. Poupard, *J. Antimicrob. Chemother.* **52**, 18 (2003).
16. Y. Hirai *et al.*, *J. Bacteriol.* **177**, 5327 (1995).
17. Y. Inamoto *et al.*, *J. Clin. Gastroenterol.* **17**, S136 (1993).
18. J. Nakayama *et al.*, *J. Biol. Chem.* **271**, 3684 (1996).
19. J. F. Tomb *et al.*, *Nature* **388**, 539 (1997).
20. H. Higashi *et al.*, *Science* **295**, 683 (2002).
21. J. B. Lowe, *Cell* **104**, 809 (2001).
22. T. O. Akama *et al.*, *Science* **295**, 124 (2002).
23. N. L. Perillo, K. E. Pace, J. J. Seilhamer, L. G. Baum, *Nature* **378**, 736 (1995).
24. D. J. Moloney *et al.*, *Nature* **406**, 369 (2000).
25. M. Demetriou, M. Granovsky, S. Quaggin, J. W. Dennis, *Nature* **409**, 733 (2001).
26. J. Nakayama, M. N. Fukuda, B. Fredette, B. Ranscht, M. Fukuda, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 7031 (1995).
27. K. Ishihara *et al.*, *Biochem. J.* **318**, 409 (1996).

28. This work was supported by a Grant-in-Aid for Scientific Research on Priority Area 14082201 from the Ministry of Education, Culture, Sports, Science and Technology of Japan (J.N.) and by grants CA 71932 (M.N.F.) and CA 33000 (M.F.) from the National Cancer Institute. The authors thank H. Ota, Y. Kawakami, T. Taketomi, and O. Harada for discussions; E. Ruoslahti, R. C. Liddington, and E. Lamar for critical reading of the manuscript; and E. Hidaka, Y. Takahashi, S. Kubota, and A. Ishida for technical assistance. This report is dedicated to the memory of Hideki Matsumoto.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/1003/DC1
Materials and Methods
Table S1
References and Notes

16 April 2004; accepted 21 June 2004

Optical Sectioning Deep Inside Live Embryos by Selective Plane Illumination Microscopy

Jan Huisken,* Jim Swoger, Filippo Del Bene, Joachim Wittbrodt, Ernst H. K. Stelzer*

Large, living biological specimens present challenges to existing optical imaging techniques because of their absorptive and scattering properties. We developed selective plane illumination microscopy (SPIM) to generate multidimensional images of samples up to a few millimeters in size. The system combines two-dimensional illumination with orthogonal camera-based detection to achieve high-resolution, optically sectioned imaging throughout the sample, with minimal photodamage and at speeds capable of capturing transient biological phenomena. We used SPIM to visualize all muscles in vivo in the transgenic Medaka line Arnie, which expresses green fluorescent protein in muscle tissue. We also demonstrate that SPIM can be applied to visualize the embryogenesis of the relatively opaque *Drosophila melanogaster* in vivo.

Modern life science research often requires multidimensional imaging of a complete spatiotemporal pattern of gene and protein expression or tracking of tissues during the development of an intact embryo (*1*). In order to visualize the precise distribution of developmental events such as activation of specific genes, a wide range of processes, from small-scale (subcellular) to large-scale (millimeters), needs to be followed. Ideally, such events, which can last from seconds to days, will be observed in live and fully intact embryos.

Several techniques have been developed that allow mapping of the three-dimensional (3D) structure of large samples (*2*). Gene expression has been monitored by in situ hybridization and block-face imaging (*3*). Techniques that provide noninvasive (optical) sectioning, as opposed to those that destroy the sample, are indispensable for live studies. Optical projection tomography can image fixed embryos at high resolution (*4*). Magnetic resonance imaging (*5*) and optical coherence tomography (*6*) feature noninvasive imaging, but do not provide specific contrasts easily.

In optical microscopy, green fluorescent protein (GFP) and its spectral variants are used for high-resolution visualization of protein localization patterns in living organisms (*7*). When GFP-labeled samples are viewed, optical sectioning (which is essential for its elimination of out-of-focus light) is obtainable by laser scanning microscopy (LSM), either by detection through a pinhole (confocal LSM) (*8*) or by exploitation of the nonlinear properties of a fluorophore (multiphoton microscopy) (*9*). Despite the improved resolution, LSM suffers from two major limitations: a limited penetration depth in heterogeneous samples and a marked difference between the lateral and axial resolution.

We developed selective plane illumination microscopy (SPIM), in which optical sectioning is achieved by illuminating the sample along a separate optical path orthogonal to the detection axis (Fig. 1 and fig. S1). A similar approach in confocal theta microscopy has been demonstrated to improve axial resolution (*10*–*12*). In SPIM, the excitation light is focused by a cylindrical lens to a sheet of light that illuminates only the focal plane of the detection optics, so that no out-of-focus fluorescence is generated (optical sectioning). The net effect is similar to that achieved by confocal LSM. However, in SPIM, only the plane currently observed is illuminated and therefore affected by bleaching. Therefore, the total number of fluorophore excitations required to image a 3D sample is greatly reduced compared to the number in confocal LSM (supporting online text).

GFP-labeled transgenic embryos of the teleost fish Medaka (*Oryzias latipes*) (*13*) were imaged with SPIM. In order to visu-

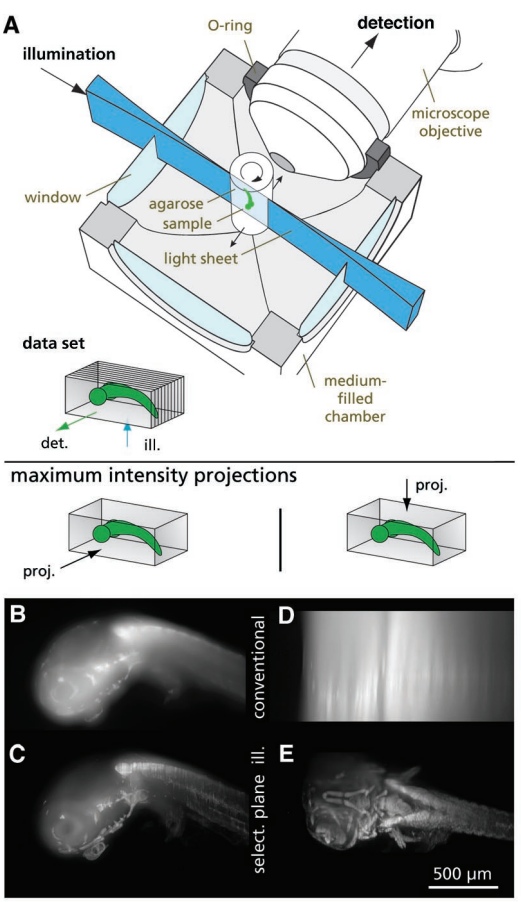


Fig. 1. (**A**) Schematic of the sample chamber. The sample is embedded in a cylinder of agarose gel. The solidified agarose is extruded from a syringe (not shown) that is held in a mechanical translation and rotation stage. The agarose cylinder is immersed in an aqueous medium that fills the chamber. The excitation light enters the chamber through a thin glass window. The microscope objective lens, which collects the fluorescence light, dips into the medium with its optical axis orthogonal to the plane of the excitation light. The objective lens is sealed with an O-ring and can be moved axially to focus on the plane of fluorescence excited by the light sheet. In a modified setup, for low-magnification lenses not corrected for water immersion, a chamber with four windows and no O-ring can be used. In this case, the objective lens images the sample from outside the chamber. det., detection; ill., illumination; proj., projection. (**B** to **E**) A Medaka embryo imaged with SPIM by two different modes of illumination. Lateral [(B) and (C)] and dorsal-ventral [(D) and (E)] maximum projections are shown. In (B) and (D), the sample was illuminated uniformly, i.e., without the cylindrical lens, as with a conventional widefield microscope. There is no optical sectioning. The elongation of fluorescent features along the detection axis is clearly visible in (D). In contrast, selective (select.) plane illumination [(C) and (E)] provided optical sectioning because the cylindrical lens focused the excitation light to a light sheet. Both image stacks were taken with a Zeiss Fluar 5×, 0.25 objective lens.

European Molecular Biology Laboratory (EMBL), Meyerhofstraβe 1, D-69117 Heidelberg, Germany.

*To whom correspondence should be addressed. E-mail: huisken@embl.de (J.H.) and stelzer@embl.de (E.H.K.S.)

alize the internal structure, we imaged the transgenic line Arnie, which expresses GFP in somatic and smooth muscles as well as in the heart (*14*). A 4-day-old fixed Arnie embryo [stage 32 (*15*)] is shown in Fig. 1. SPIM was capable of resolving the internal structures of the entire organism with high resolution (better than 6 μm) as deep as 500 μm inside the fish, a penetration depth that cannot be reached using confocal LSM (fig. S6). The axial resolution in SPIM is determined by the lateral width of the light sheet; for the configuration shown in Fig. 1, the axial extent of the point spread function (PSF) was about 6 μm, whereas without the light sheet it was more than 20 μm (supporting online text).

Any fluorescence imaging system suffers from scattering and absorption in the tissue; in large and highly scattering samples, the image quality decreases as the optical path length in the sample increases. This problem can be reduced by a multiview reconstruction, in which multiple 3D data sets of the same object are collected from different directions and combined in a postprocessing step (*16*–*18*). The high-quality information is extracted from each data set and merged into a single, superior 3D image (supporting online text). One way to do this is by parallel image acquisition, using more than one lens for the detection of fluorescence (*18*).

We collected SPIM data for a multiview reconstruction sequentially by generating multiple image stacks between which the sample was rotated. Sample deformations were avoided with a rotation axis parallel to gravity (Fig. 1). In contrast to tomographic reconstruction techniques [such as those in (*4*)], which require extensive processing of the data to yield any meaningful 3D information, rotation and subsequent data processing are optional in SPIM. They allow a further increase in image quality and axial resolution compared to a single stack, but in many cases a single, unprocessed 3D SPIM stack alone provides sufficient information.

We performed a multiview reconstruction with four stacks taken with four orientations of the same sample (figs. S2 and S3). Combination of these stacks (supporting online text) yielded a complete view of the sample, ~1.5 mm long and ~0.9 mm wide. In Fig. 2, the complete fused data set is shown and the most pronounced tissues are labeled. The decrease in image quality with penetration depth is corrected by the fusion process. It yielded an increased information content in regions that were obscured (by absorption or scattering in the sample) in some of the unprocessed single views.

The method of embedding the sample in a low-concentration agarose cylinder is nondisruptive and easily applied to live embryos. We routinely image live Medaka and *Drosophila* embryos over periods of up to 3 days without detrimental effects on embryogenesis and development. To demonstrate the potential of SPIM technology, we investigated the Medaka heart, a structure barely accessible by conventional confocal LSM because of its ventral position in the yolk cell. We imaged transgenic Medaka Arnie embryos and show a reconstruction of the inner heart surface (Fig. 3A) derived from the data set shown in Fig. 2. This reveals the internal structure of the heart ventricle and atrium. In a slightly earlier stage, internal organs such as the heart and other mesoendodermal deriva-

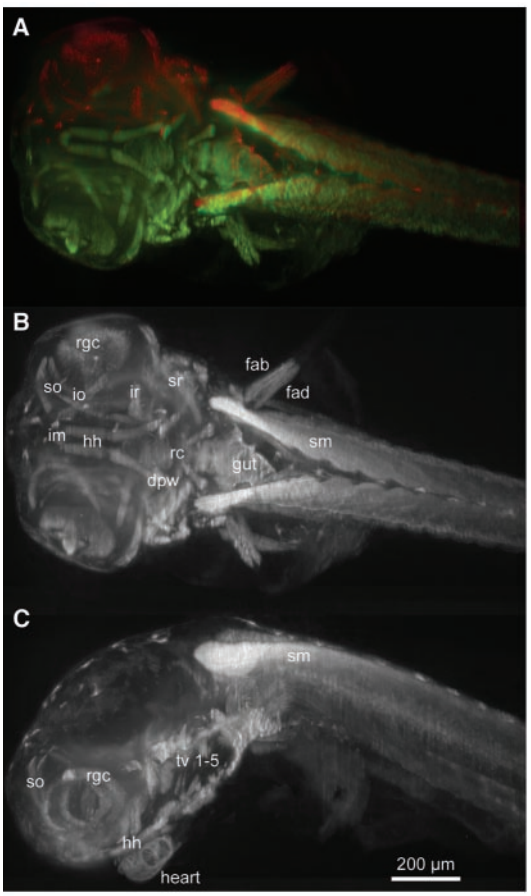


Fig. 2. A Medaka embryo (the same as in Fig. 1) imaged with SPIM and processed by multiview reconstruction (figs. S2, S3, and S6 and movies S1 and S2). (**A**) Overlay of a single stack (green) and the fusion of four data sets (red and green). (**B**) Dorsal-ventral and (**C**) lateral maximum intensity projections of the fused data. The high resolution throughout the entire fish allows identification of different tissues: rgc, retinal ganglion cells; so, superior oblique; io, inferior oblique; ir, inferior rectus; sr, superior rectus; im, intermandibualaris; hh, hyohyal; rc, rectus communis; dpw, dorsal pharyngeal wall; fad, fin adductor; fab, fin abductor; sm, somitic mesoderm; tv, transverse ventrals. The stack has a size of 1201 by 659 by 688 pixels (1549 μm by 850 μm by 888 μm).

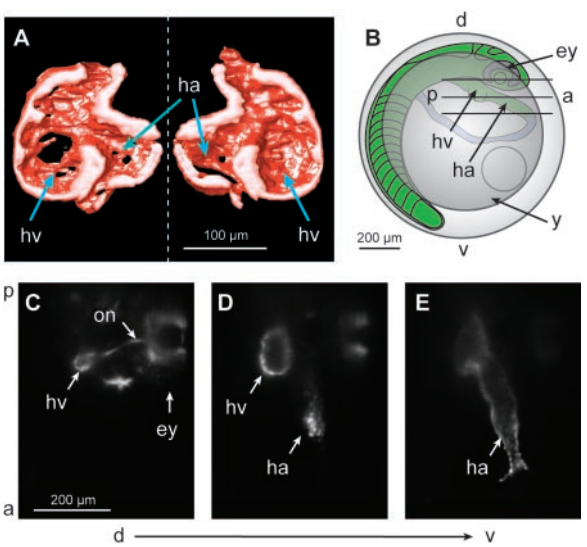


Fig. 3. A Medaka heart imaged with SPIM (movies S3 and S4). (**A**) Surface rendering of the heart taken from the data shown in Fig. 2. The heart has been cut open computationally to make the internal structure visible. hv, heart ventricle; ha, heart atrium. (**B**) Schematic representation of a Medaka embryo at stage 26 of development (*13*), 2 days post-fertilization. Three optically sectioned planes are indicated. At this stage, ventral structures such as the heart are deeply buried in the yolk sphere. d, dorsal; v, ventral; a, anterior; p, posterior; y, yolk; ey, eye. (**C**) Optical section of an Arnie embryo showing the eye and the optic nerve labeling and the dorsal part of the heart ventricle. on, optic nerve. (**D**) Optical section showing the heart ventricle chamber and the dorsal wall of the heart atrium. (**E**) Optical section showing the atrium chamber.

tives are deeply buried in the yolk sphere, under the body of the embryo (Fig. 3B). In Fig. 3, C to E, three optical sections at different depths illustrate GFP expression in the muscles of the living heart. Fast frame recording (10 frames per s) allows imaging of the heartbeat (movies S3 and S4); similar imaging has previously only been demonstrated at stages when the heart is exposed and by cooling the embryo to reduce the heart rate (*19*).

To demonstrate that SPIM can also be used to image the internal structures of relatively opaque embryos, we recorded a time series (movie S5) of the embryogenesis of the fruit fly *Drosophila melanogaster* (Fig. 4). GFP-moesin labeled the plasma membrane throughout the embryo (*20*). Even without multiview reconstruction, structures inside the embryo are clearly identifiable and traceable. Stacks (56 planes each) were taken automatically every 5 min over a period of 17 hours, without refocusing or realignment. Even after being irradiated for 11,480 images, the embryo was unaffected and completed embryogenesis normally.

In summary, we present an optical widefield microscope capable of imaging protein expression patterns deep inside both fixed and live embryos. By selective illumination of a single plane, the excitation light is used efficiently to achieve optical sectioning and reduced photodamage in large samples, key features in the study of embryonic development. The method of sample mounting allows positioning and rotation to orient the sample for optimal imaging conditions. The optional multiview reconstruction combines independently acquired data sets into an optimal representation of the sample. The implementation of other contrasts such as scattered light will be straightforward. The system is compact, fast, optically stable, and easy to use.

SPIM is well suited for the visualization of high-resolution gene and protein expression patterns in three dimensions in the context of morphogenesis. Heart function and development can be precisely followed in vivo using SPIM in Arnie transgenic embryos. Because of its speed and its automatable operation, SPIM can serve as a tool for large-scale studies of developing organisms and the systematic and comprehensive acquisition and collection of expression data. Even screens for molecules that interfere with development and regeneration on a medium-throughput scale seem feasible. SPIM technology can be readily applied to a wide range of organisms, from whole embryos to single cells. Subcellular resolution can be obtained in live samples kept in a biologically relevant environment within the organism or in culture. Therefore, SPIM also has the potential to be of use in the promising fields of 3D cultured cells (*21*) and 3D cell migration (*22*).

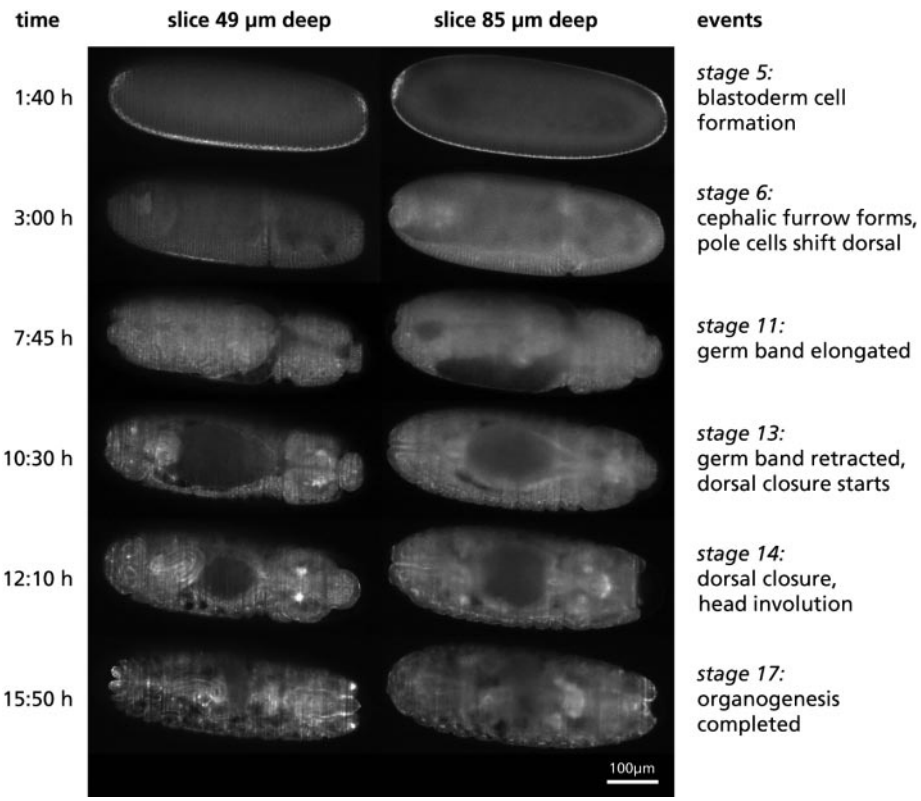


Fig. 4. Time-lapse imaging of *Drosophila melanogaster* embryogenesis. Six out of 205 time points acquired are shown (movie S5). At each time point, 56 planes were recorded, from which two (at depths of 49 μm and 85 μm below the cortex) are shown. No multiview reconstruction was necessary. The optical sectioning capability and the good lateral resolution are apparent. Despite the optically dense structure of the *Drosophila* embryo, features are well resolved at these depths in the sample. For this figure, the images were oriented so that the illumination occurs from below. This results in a slight drop in intensity and clarity from the bottom to the top of each slice. Nevertheless, the information content across the embryo is nearly uniform, and the overall morphogenetic movements during embryonic development can be followed. The images were normalized to exhibit the same overall intensity, thus compensating the continuous production of GFP-moesin. We took 205 stacks at 5-min intervals with a Zeiss Achroplan 10×, 0.30W objective lens (56 planes per stack at 4-μm spacing) for 11,480 images in total.

References and Notes

1. S. G. Megason, S. E. Fraser, *Mech. Dev.* **120**, 1407 (2003).
2. S. W. Ruffins, R. E. Jacobs, S. E. Fraser, *Curr. Opin. Neurobiol.* **12**, 580 (2002).
3. W. J. Weninger, T. Mohun, *Nature Genet.* **30**, 59 (2002).
4. J. Sharpe *et al.*, *Science* **296**, 541 (2002).
5. A. Y. Louie *et al.*, *Nature Biotechnol.* **18**, 321 (2000).
6. D. Huang *et al.*, *Science* **254**, 1178 (1991).
7. M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, D. C. Prasher, *Science* **263**, 802 (1994).
8. J. B. Pawley, *Handbook of Biological Confocal Microscopy* (Plenum, New York, 1995).
9. W. Denk, J. H. Strickler, W. W. Webb, *Science* **248**, 73 (1990).
10. E. H. K. Stelzer, S. Lindek, *Opt. Commun.* **111**, 536 (1994).
11. E. H. K. Stelzer *et al.*, *J. Microsc.* **179**, 1 (1995).
12. S. Lindek, J. Swoger, E. H. K. Stelzer, *J. Mod. Opt.* **46**, 843 (1999).
13. J. Wittbrodt, A. Shima, M. Schartl, *Nature Rev. Genet.* **3**, 53 (2002).
14. Materials and methods are available as supporting material on *Science* Online.
15. T. Iwamatsu, *Zool. Sci.* **11**, 825 (1994).
16. P. J. Shaw, *J. Microsc.* **158**, 165 (1990).
17. S. Kikuchi, K. Sonobe, S. Mashiko, Y. Hiraoka, N. Ohyama, *Opt. Commun.* **138**, 21 (1997).
18. J. Swoger, J. Huisken, E. H. K. Stelzer, *Opt. Lett.* **28**, 1654 (2003).
19. J. R. Hove *et al.*, *Nature* **421**, 172 (2003).
20. K. A. Edwards, M. Demsky, R. A. Montague, N. Weymouth, D. P. Kiehart, *Dev. Biol.* **191**, 103 (1997).
21. A. Abbott, *Nature* **424**, 870 (2003).
22. D. J. Webb, A. F. Horowitz, *Nature Cell Biol.* **5**, 690 (2003).

23. We thank S. Enders and K. Greger for contributions to the instrumentation and F. Jankovics and D. Brunner for providing the *Drosophila* samples. The beating-heart data was recorded by K. Greger.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/1007/DC1
Materials and Methods
SOM Text
Figs. S1 to S6
References and Notes
Movies S1 to S5

6 May 2004; accepted 15 July 2004

Increased Nuclear NAD Biosynthesis and SIRT1 Activation Prevent Axonal Degeneration

Toshiyuki Araki, Yo Sasaki, Jeffrey Milbrandt*

Axonal degeneration is an active program of self-destruction that is observed in many physiological and pathological settings. In Wallerian degeneration slow (*wld*^{s}) mice, Wallerian degeneration in response to axonal injury is delayed because of a mutation that results in overexpression of a chimeric protein (Wld^s) composed of the ubiquitin assembly protein Ufd2a and the nicotinamide adenine dinucleotide (NAD) biosynthetic enzyme Nmnat1. We demonstrate that increased Nmnat activity is responsible for the axon-sparing activity of the Wld^s protein. Furthermore, we demonstrate that SIRT1, a mammalian ortholog of Sir2, is the downstream effector of increased Nmnat activity that leads to axonal protection. These findings suggest that novel therapeutic strategies directed at increasing the supply of NAD and/or Sir2 activation may be effective for treatment of diseases characterized by axonopathy and neurodegeneration.

Axonopathy is a critical feature of many peripheral neuropathies, and axonal degeneration often precedes the death of neuronal cell bodies in neurodegenerative diseases such as Parkinson's and Alzheimer's disease (*1*). These axonal deficits are an important component of the patient's disability and potentially represent a therapeutic target for combating these diseases (*2*).

The discovery of a spontaneous dominant mutation in mice that results in delayed axonal degeneration, the Wallerian degeneration slow (*wld*^{s}) mice, suggests that axonal degeneration is an active process of self-destruction (*3*). Genetic analysis has shown that the *wld*^{s} mutation comprises an 85-kb tandem triplication, which results in overexpression of a chimeric nuclear molecule (Wld^s protein). This protein is composed of the N-terminal 70 amino acids of Ufd2a (ubiquitin fusion degradation protein 2a), a ubiquitin-chain assembly factor, fused to the complete sequence of nicotinamide mononucleotide adenylyltransferase1 (Nmnat1), an enzyme in the NAD biosynthetic pathway that generates NAD within the nucleus (*4*, *5*). The Wld^s protein has Nmnat activity but lacks ubiquitin ligase function, suggesting that axonal protection is derived from either increased Nmnat1 activity or a "dominant negative" inhibition of Ufd2a function.

To determine the mechanism of delayed axonal degeneration mediated by the Wld^s protein, we used an in vitro Wallerian degeneration model. Primary dorsal root ganglion (DRG) explant neurons were infected with lentivirus expressing recombinant proteins, and axons were injured by either removal of the neuronal cell body (transection) or growth in vincristine (toxic). We first demonstrated that transected axons from neurons expressing the Wld^s protein degenerated with the delayed kinetics characteristic of neurons derived from *wld*^{s} mice (Fig. 1A) (*6*). Next, we compared axonal degeneration after transection in wild-type neurons that express the chimeric Wld^s protein with those that express only the Ufd2a or Nmnat1 portions of the Wld^s protein, linked to enhanced green fluorescent protein (EGFP) (Fig. 1B). We found that expression of EGFP-Nmnat1 delayed axonal degeneration comparable to Wld^s protein itself, whereas the N-terminal 70 amino acids of Ufd2a (fused to EGFP), targeted to either the nucleus or cytoplasm, did not affect axonal degeneration. Quantification of these effects was performed by counting the percentage of remaining neurites at various times after removal of neuronal cell bodies. This analysis showed that EGFP-Nmnat1, like Wld^s protein itself, resulted in a >10-fold increase in intact neurites 72 hours after injury. To further exclude direct involvement of the ubiquitin-proteasome system in Wld^s protein-mediated axonal protection, we examined the effect of Ufd2a inhibition using either a dominant-negative Ufd2a mutant or a Ufd2a small interfering RNA (siRNA) construct. However, neither method of Ufd2a inhibition resulted in delayed axonal degradation in response to axotomy. Together, these experiments demonstrated that the Nmnat1 portion of the Wld^s protein is responsible for the delayed axonal degeneration observed in *wld*^{s} mice.

Nmnat1 is an enzyme in the nuclear NAD biosynthetic pathway that catalyzes the conversion of nicotinamide mononucleotide (NMN) and nicotinate mononucleotide (NaMN) to NAD and nicotinate adenine dinucleotide (NaAD), respectively (*7*). The axonal protection observed in Nmnat1 overexpressing neurons could be mediated by its ability to synthesize NAD (i.e., its enzymatic activity), or perhaps by other unknown functions of this protein. To address this question, we used the Nmnat1 crystal structure to identify several residues predicted to participate in substrate binding (*8*). A mutation in one of these residues was engineered into full length Nmnat1 (W170A) and Wld^s (W258A) protein. In vitro enzymatic assays confirmed that both of these mutant proteins were severely limited in their ability to synthesize NAD (fig. S2). Each of these mutants and their respective wild-type counterparts was singly introduced into neurons to assess their ability to protect axons from degradation. We found that neurons expressing these enzymatically inactive mutants had no axonal protective effects (Fig. 1C), which indicates that NAD/NaAD production is responsible for the ability of Nmnat1 to prevent axonal degradation.

In addition to mechanical transection, axonal protection in *wld*^{s} mice is also observed against other damaging agents such as ischemia and toxins (*2*, *9*). We sought to determine whether increased Nmnat activity would also delay axonal degradation in response to other types of axonal injury, such as vincristine, a cancer chemotherapeutic reagent with well-characterized axonal toxicity. Neurons expressing either Nmnat1 or EGFP (control) were grown in 0.5 μM vincristine for up to 9 days. We found that axons of neurons expressing Nmnat1 maintained their original length and refractility, whereas axons emanating from uninfected neurons or those expressing EGFP gradually retracted and had mostly degenerated by day 9 (Fig. 2). These results indicate that increased Nmnat activity itself can protect axons from both mechanical and toxic insults.

Previous experiments have shown that neuronal cells express membrane proteins that can bind and transport extracellular NAD into the cell (*10*). This encouraged us to investigate whether exogenously administered NAD could prevent axonal degeneration. We added various concentrations of NAD to neuronal cultures before axonal transection and examined the extent of axonal degradation. We found that 0.1 to 1 mM NAD added 24 hours before axotomy significantly delayed axonal degeneration, although exogenously applied NAD (1 mM) was slightly less effective in protecting axons than lentivirus-mediated Nmnat1 expression (Fig. 3A). These results provide direct support for the idea

Department of Pathology, Washington University School of Medicine, St. Louis, Missouri 63110, USA.

*To whom correspondence should be addressed. E-mail: jmilbrandt@wustl.edu

that increased NAD supply can prevent axonal degradation.

NAD plays a variety of roles in the cell. In the mitochondria, it is involved in electron-transport processes important in energy metabolism, whereas in the nucleus NAD regulates aspects of DNA repair and transcription. In yeast, the Nmnat homologs are nuclear proteins that participate in the nuclear NAD salvage pathway (*11*, *12*), which suggests that NAD could be mediating its axonal protective effects by a nuclear mechanism. Indeed, both Wld^s and Nmnat1 were found in the nucleus with immunohistochemistry and EGFP fluorescence (fig. S3). Interestingly, the activation of the NAD salvage pathway in yeast does not alter total cellular NAD levels (*11*). Similarly, tissue NAD levels in wild-type and *wld^s* brain are similar, despite the increased NAD synthetic activity in *wld^s* tissues (*5*). We measured NAD levels in wild-type and Nmnat1-expressing cells using sensitive microscale enzymatic assays (*13*) and found that increased Nmnat activity did not result in changes in overall cellular NAD levels (*14*). Together, these data suggest that an NAD-dependent enzymatic activity in the nucleus, as opposed to cytoplasmic NAD-dependent processes, is likely to mediate the axonal protection observed in response to increased Nmnat activity.

To gain further insight into the mechanism of NAD-dependent axonal protection (NDAP), we examined whether NAD was required prior to the removal of the neuronal cell bodies or whether direct exposure of the severed axons to high levels of NAD was sufficient to provide protection (Fig. 3B). Neuronal cultures were prepared, and 1 mM NAD was added to the culture medium at the time of axonal transection or at various times (4 to 48 hours) before injury. We found that administering NAD at the time of axonal transection or for up to 8 hours before injury had no protective effects on axons. However, significant axon sparing was observed when neurons were incubated with NAD for longer periods of time before injury, with the greatest effects occurring after 24 hours of NAD pretreatment. These results indicate that NDAP is not mediated by a rapid posttranslational modification within the axons themselves. Instead, they suggest that the protective process requires de novo transcriptional and/or translational events. The active nature of axonal self-destruction was further emphasized by our observations that treatment of neurons for 24 hours before axotomy with inhibitors of either RNA (actinomycin D) or protein (cycloheximide) synthesis resulted in axonal protection (*15*).

The Sir2 family of protein deacetylases and poly(ADP-ribose) polymerase (PARP) are involved in major NAD-dependent nuclear enzymatic activities. Sir2 is an NAD-

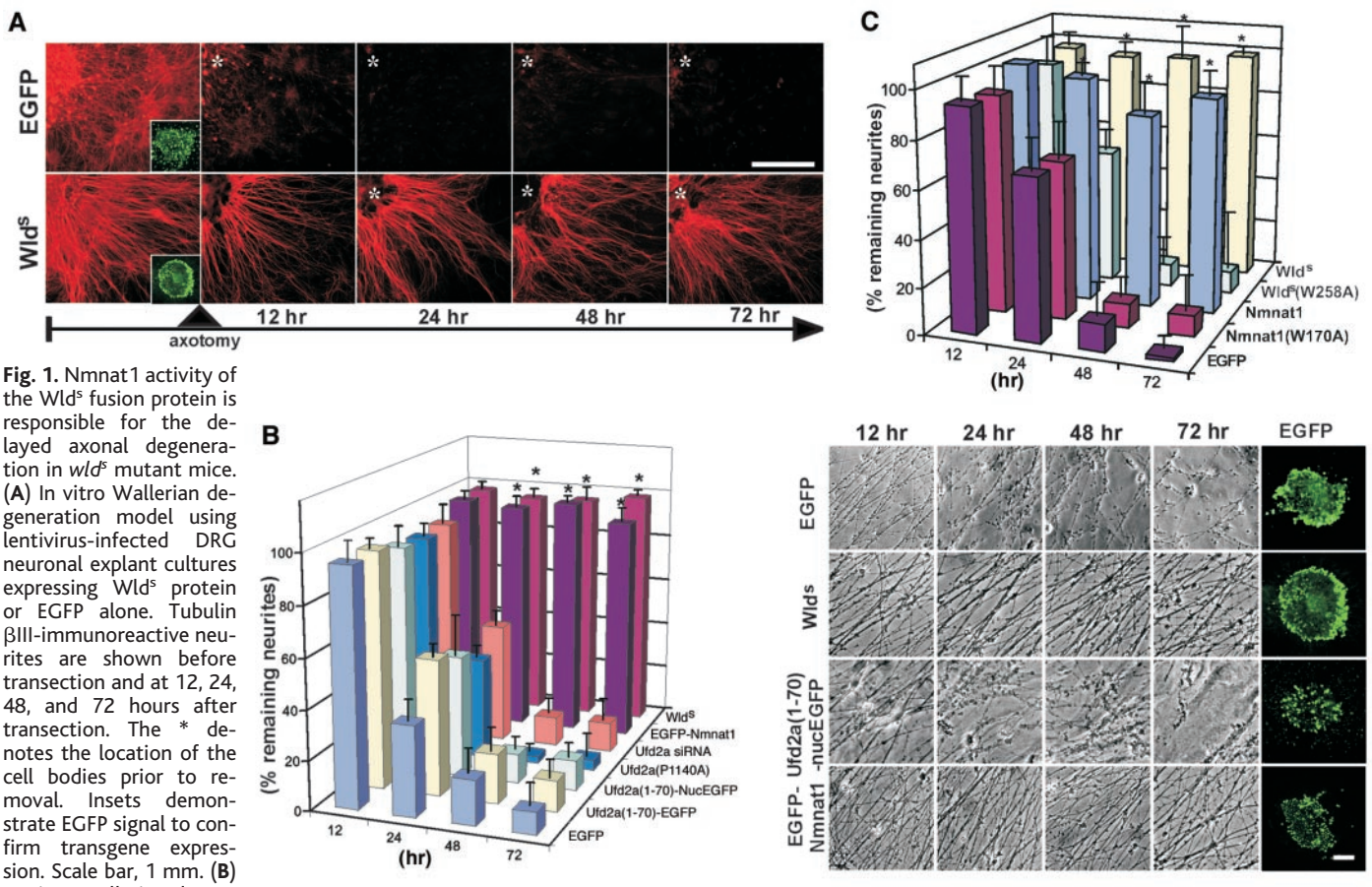


Fig. 1. Nmnat1 activity of the Wld^s fusion protein is responsible for the delayed axonal degeneration in *wld^s* mutant mice. (**A**) In vitro Wallerian degeneration model using lentivirus-infected DRG neuronal explant cultures expressing Wld^s protein or EGFP alone. Tubulin βIII-immunoreactive neurites are shown before transection and at 12, 24, 48, and 72 hours after transection. The * denotes the location of the cell bodies prior to removal. Insets demonstrate EGFP signal to confirm transgene expression. Scale bar, 1 mm. (**B**) In vitro Wallerian degeneration model with lentivirus-infected DRG neurons expressing EGFP only, Wld^s protein, Ufd2a portion (70 residues) of Wld^s protein fused to EGFP [Ufd2a(1-70)-EGFP], Ufd2a(1-70)-EGFP with C-terminal nuclear localization signal, Nmnat1 portion of Wld^s protein fused to EGFP, dominant-negative Ufd2a [Ufd2a(P1140A)], or Ufd2a siRNA construct. Representative images of neurites and quantitative analysis of remaining neurite numbers (percentage of remaining neurites relative to pretransection ± SD) at the indicated time point with each construct are shown. The * indicates significant difference ($P < 0.0001$) with EGFP-infected neurons. The EGFP signal before transection confirms transgene expression (bottom row). Scale bar, 50 μm. (**C**) In vitro Wallerian degeneration model of lentivirus-infected DRG neurons expressing Nmnat1 or Wld^s protein, mutants of these proteins that lack NAD-synthesis activity Nmnat1(W170A) and Wld^s(W258A), or EGFP (see color code). Quantitative analysis of the number of remaining neurites at the indicated time points for each construct (percentage of remaining neurites relative to pretransection ± SD). The * indicates significant difference ($P < 0.0001$) with EGFP-infected neurons.

dependent deacetylase of histones (*15*) and other proteins, and its activation is central to promoting increased longevity in yeast and *Caenorhabditis elegans* (*17*, *18*). PARP is activated by DNA damage and is involved in DNA repair (*19*). The importance of these NAD-dependent enzymes in regulating gene activity prompted us to investigate their potential role in the self-destructive process of axonal degradation. We tested whether inhibitors of Sir2 (Sirtinol) (*20*) and PARP [3-aminobenzamide (3AB)] (*21*) could affect NDAP (Fig. 4A). Neurons were cultured in the presence of 1 mM NAD and either Sirtinol (100 μM) or 3AB (20 mM). Axonal transection was performed by removal of the neuronal cell bodies, and the extent of axonal degradation was assessed 12 to 72 hours later. We found that, although Sirtinol had no axonal toxicity on uninjured axons (fig. S5), it effectively blocked NDAP after transection, indicating that Sir2 proteins are likely effectors of this process. In contrast, 3AB had no effect on NDAP, indicating that PARP does not play a role in axonal protection. Interestingly, 3AB alone did stimulate limited axonal protection (Fig. 4A), presumably as a consequence of PARP inhibition, which decreases NAD consumption and raises nuclear NAD levels. To confirm the involvement of Sir2 proteins in NDAP, we tested the effects of resveratrol (10 to 100 μM), a polyphenol compound found in grapes that enhances Sir2 activity (*22*). We found that neurons treated with resveratrol prior to axotomy showed a decrease in axonal degradation that was comparable to that obtained with NAD (Fig. 4B), providing further support for the idea that Sir2 proteins are effectors of the axonal protection mediated by increased Nmnat activity.

In humans and rodents, seven molecules that share the Sir2 conserved domain [sirtuin (SIRT) 1 to 7] have been identified (*23*). SIRT1 is located in the nucleus and is involved in chromatin remodeling and the regulation of transcription factors such as p53 (*24*), whereas other SIRT proteins are located within the cytoplasm and mitochondria (*25*, *26*). To determine which SIRT protein(s) is involved in NDAP, we performed knockdown experiments using siRNA constructs to specifically target each member of the SIRT family. Neurons were infected with lentiviruses expressing specific SIRT siRNA constructs that effectively suppressed expression of their intended target (table S1). The infected neurons were cultured in 1 mM NAD, and axonal transection was performed by removing the cell bodies. Inhibiting the expression of most SIRT proteins did not significantly affect NDAP; however, the knockdown of SIRT1 blocked NDAP as effectively as Sirtinol (Fig. 4C). Like Sirtinol treatment, SIRT1 inhibition by siRNA did not affect the rate of degeneration in untreated neurons or the axonal integrity in uninjured neurons (fig. S5). These results indicate that SIRT1 is the major effector of the increased NAD supply that effectively prevents axonal self-destruction. Although, SIRT1 may deacetylate proteins directly involved in axonal stability, its predomi-

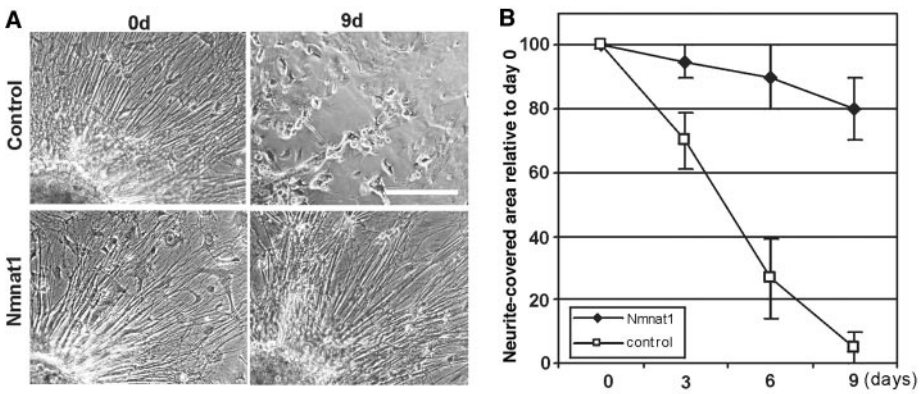


Fig. 2. Increased Nmnat1 activity protects axons from degeneration caused by vincristine toxicity. (**A**) DRG neuronal explants expressing either Nmnat1 or EGFP (control) were cultured with 0.5 μM vincristine. Representative images of neurites (phase-contrast) at the indicated times after vincristine addition are shown. Scale bar, 1 mm. (**B**) Quantification of the protective effect at the indicated time points is plotted as the area covered by neurites relative to that covered by neurites before treatment.

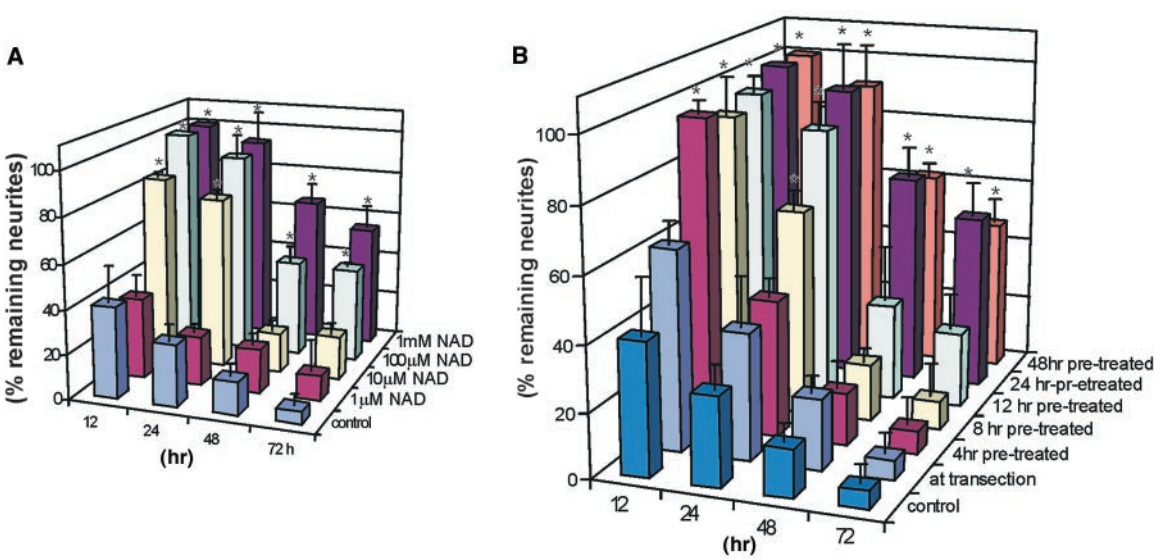


Fig. 3. Axonal protection requires pretreatment of neurons with NAD before injury. (**A**) In vitro Wallerian degeneration using DRG explants cultured in the presence of various concentrations of NAD added 24 hours before axonal transection. (**B**) DRG explants preincubated with 1mM NAD for 4, 8, 12, 24, or 48 hours prior to transection. In each experiment, the number of remaining neurites (percentage of remaining neurites relative to pretransection ± SD) is shown at each of the indicated time points. The * indicates significant axonal protection compared with control ($P < 0.0001$).

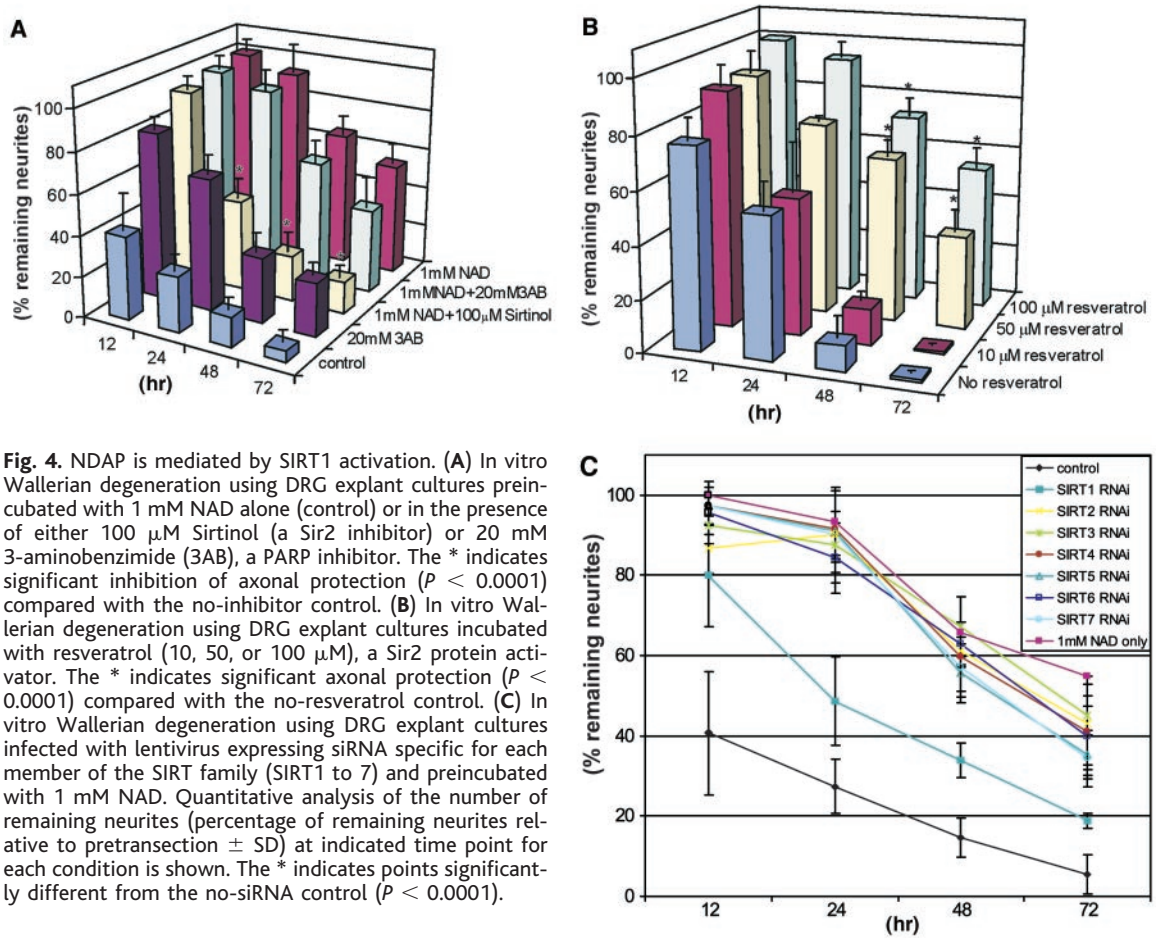


Fig. 4. NDAP is mediated by SIRT1 activation. (**A**) In vitro Wallerian degeneration using DRG explant cultures preincubated with 1 mM NAD alone (control) or in the presence of either 100 μM Sirtinol (a Sir2 inhibitor) or 20 mM 3-aminobenzimide (3AB), a PARP inhibitor. The * indicates significant inhibition of axonal protection ($P < 0.0001$) compared with the no-inhibitor control. (**B**) In vitro Wallerian degeneration using DRG explant cultures incubated with resveratrol (10, 50, or 100 μM), a Sir2 protein activator. The * indicates significant axonal protection ($P < 0.0001$) compared with the no-resveratrol control. (**C**) In vitro Wallerian degeneration using DRG explant cultures infected with lentivirus expressing siRNA specific for each member of the SIRT family (SIRT1 to 7) and preincubated with 1 mM NAD. Quantitative analysis of the number of remaining neurites (percentage of remaining neurites relative to pretransection ± SD) at indicated time point for each condition is shown. The * indicates points significantly different from the no-siRNA control ($P < 0.0001$).

nantly nuclear location, along with the requirement for NAD ~24 hours prior to injury for effective protection, suggest that SIRT1 regulates a genetic program that leads to axonal protection.

In *wld*^{s} mice, axonal protection through Wld^s protein overexpression has been demonstrated in models of motor neuron and Parkinson's disease and in peripheral sensory neurons affected by chemotherapeutic agents (*2*, *27*). Our results indicate that the molecular mechanism of axonal protection in the *wld*^{s} mice is due to the increased nuclear NAD biosynthesis that results from increased Nmnat1 activity and consequent activation of the protein deacetylase SIRT1. Other intracellular events that affect NAD levels or NAD/NADH ratios, such as energy production through respiration, may also affect physiological and pathological processes in the nervous system through SIRT1-dependent pathways (*28*). It is possible that the alteration of NAD levels by manipulation of the NAD biosynthetic pathway, Sir2 protein activity, or other downstream effectors will provide new therapeutic opportunities for the treatment of diseases involving axonopathy and neurodegeneration.

References and Notes

1. M. C. Raff, A. V. Whitmore, J. T. Finn, *Science* **296**, 868 (2002).
2. M. P. Coleman, V. H. Perry, *Trends Neurosci.* **25**, 532.
3. E. R. Lunn *et al.*, *Eur. J. Neurosci.* **1**, 27 (1989).
4. L. Conforti *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11377 (2000).
5. T. G. Mack *et al.*, *Nat. Neurosci.* **4**, 1199 (2001).
6. E. A. Buckmaster, V. H. Perry, M. C. Brown, *Eur. J. Neurosci.* **7**, 1596 (1995).
7. G. Magni *et al.*, *Cell. Mol. Life Sci.* **61**, 19 (2004).
8. T. Zhou *et al.*, *J. Biol. Chem.* **277**, 13148 (2002).
9. T. H. Gillingwater *et al.*, *J. Cereb. Blood Flow Metab.* **24**, 62 (2004).
10. S. Bruzzone *et al.*, *FASEB J.* **15**, 10 (2001).
11. R. M. Anderson *et al.*, *J. Biol. Chem.* **277**, 18881 (2002).
12. W. K. Huh *et al.*, *Nature* **425**, 686 (2003).
13. C. Szabo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1753 (1996).
14. T. Araki *et al.*, unpublished observations.
15. T. Araki *et al.*, unpublished observations.
16. S. Imai *et al.*, *Nature* **403**, 795 (2000).
17. M. Kaeberlein, M. McVey, L. Guarente, *Genes Dev.* **13**, 2570 (1999).
18. H. A. Tissenbaum, L. Guarente, *Nature* **410**, 227 (2001).
19. S. D. Skaper, *Ann. N.Y. Acad. Sci.* **993**, 217, discussion 287 (2003).
20. C. M. Grozinger *et al.*, *J. Biol. Chem.* **276**, 38837 (2001).
21. G. J. Southan, C. Szabo, *Curr. Med. Chem.* **10**, 321 (2003).
22. K. T. Howitz *et al.*, *Nature* **425**, 191 (2003).
23. S. W. Buck, C. M. Gallo, J. S. Smith, *J. Leukoc. Biol.* **75**, 939, 2004.
24. J. Luo *et al.*, *Cell* **107**, 137 (2001).
25. P. Onyango *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 13653 (2002).
26. B. J. North *et al.*, *Mol. Cell* **11**, 437 (2003).
27. A. Sajadi, B. L. Schneider, P. Aebischer, *Curr. Biol.* **14**, 326 (2004).
28. S. J. Lin *et al.*, *Genes Dev.* **18**, 12 (2004).
29. We thank D. Baltimore for the lentiviral expression system, Kazusa DNA Research Institute for the murine Ufd2a cDNA, and J. Manchester and J. Gordon for assistance with NAD measurements. We thank members of the laboratory and our colleagues E. Johnson, J. Gordon, S. Imai, R. Van Gelder, and R. Heuckeroth for helpful discussion and comments on the manuscript. The work was supported by grants from the National Institute of Neurological Disorders and Stroke NS40745 and National Institute on Aging AG13730, and a pilot grant from the Alzheimer's Disease Research Center at Washington University (National Institute on Aging AG05681).

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/1010/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Table S1
References

17 March 2004; accepted 29 June 2004

Evidence for Addiction-like Behavior in the Rat

Véronique Deroche-Gamonet, David Belin, Pier Vincenzo Piazza*

Although the voluntary intake of drugs of abuse is a behavior largely preserved throughout phylogeny, it is currently unclear whether pathological drug use ("addiction") can be observed in species other than humans. Here, we report that behaviors that resemble three of the essential diagnostic criteria for addiction appear over time in rats trained to self-administer cocaine. As in humans, this addiction-like behavior is present only in a small proportion of subjects using cocaine and is highly predictive of relapse after withdrawal. These findings provide a new basis for developing a true understanding and treatment of addiction.

The voluntary intake of drugs of abuse is a behavior largely conserved throughout phylogeny. Preferences for drug-associated environments or drug-reinforced learning of tasks have been found in several species (*1–6*). The possibility of studying these behaviors in animals has helped us to understand the neurobiological basis of drug taking (*7–10*) and, more generally, the brain systems for reward (*11*).

As important as the comprehension of drug taking and reward is, however, the major goal of drug abuse research is to uncover the mechanisms of addiction. Addiction is not just the taking of drugs but compulsive drug use maintained despite adverse consequences for the user (*12*). This pathological behavior appears only in a small proportion (15 to 17%) of those using drugs (*13*) and has the characteristics of a chronic disease (*12*). Indeed, even after a prolonged period of withdrawal, 90% of addicted individuals relapse to drug taking (*14*). Unfortunately, our knowledge of the biological basis of addiction lags behind our knowledge of the mechanisms of drug taking, probably because convincing evidence of addiction in animals is lacking.

We thus investigated whether addiction-like behaviors can be observed in rodents. Our experiments used intravenous self-administration (SA), the most common procedure for the study of voluntary drug intake in laboratory animals. Freely moving rats learned to obtain intravenous infusions of cocaine by poking their noses into a hole. To allow for addiction-like behavior to appear, we studied SA over a time frame of about 3 months, much longer than is typical in SA experiments (i.e., between 10 and 30 days). During this prolonged SA period, we repeatedly evaluated the intensity of three behaviors resembling those currently considered the hallmarks of substance dependence in the DSM-IV (*12*):

(i) The subject has difficulty stopping drug use or limiting drug intake. We measured the persistence of cocaine seeking during a period of signaled nonavailability of cocaine. The daily SA session included three 40-min "drug periods" that were separated by two 15-min "no-drug periods." During the drug periods, a standard FR5 reinforcement schedule was in effect: Five nose-pokes resulted in an infusion of 0.8 mg of cocaine per kilogram of body weight (mg/kg). During the no-drug periods, nose-pokes had no effect. The two different periods of drug availability were signaled by a change in the illumination of the SA chamber (*15*).

(ii) The subject has an extremely high motivation to take the drug, with activities focused on its procurement and consumption. We used a progressive-ratio schedule: The number of responses required to receive one infusion of cocaine (i.e., the ratio of responding to reward) was increased progressively within the SA session. The maximal amount of work that the animal will perform before cessation of responding, referred to as the breaking point, is considered a reliable index of the motivation for the drug (*16*).

(iii) Substance use is continued despite its harmful consequences. We measured the persistence of the animals' responding for the drug when drug delivery was associated with a punishment. During these sessions, nose-pokes on the standard FR5 schedule resulted in the delivery of both the drug and an electric shock. This shock punishment was signaled by a new cue light that was turned on at the time of the first nose-poke and off after the delivery of the shock (*15*).

To provide further validity to the addiction-like behaviors studied here, we analyzed their development as a function of the propensity of an individual to relapse to drug seeking. This approach was chosen because, as mentioned above, in humans the most predictable outcome of a first diagnosis of addiction is a 90% chance of relapse to drug use even after long periods of withdrawal (*14*). To study the propensity to relapse, we used the "reinstatement" procedure (*17*). After a 5- or 30-day period of withdrawal that followed the 3 months of SA, rats were exposed to stimuli known to induce relapse in humans, such as small amounts of the abused drug or a conditioned stimulus associated with drug taking. These challenges induce high levels of responding (reinstatement) on the device previously associated with drug delivery. The rate of responding during the test for reinstatement is considered a measure of the propensity to relapse.

In a first experiment, rats ($n = 17$) were assigned to two groups on the basis of their behavior on the test for reinstatement, induced here by the infusion of small quantities of cocaine given after 5 days of withdrawal that followed 76 days of testing for SA (*15*). The two groups ($n = 7$ each) contained the rats with the 40% highest (HRein) and 40% lowest (LRein) cocaine-induced reinstatement of responding (Fig. 1D) (*15*). HRein and LRein differed profoundly on the occurrence of addiction-like behaviors (Fig. 1, A to C) (*15*). HRein rats progressively increased their drug-seeking behavior during the no-drug periods ($F_{2,12} = 3.54$, $P < 0.05$) and after punishment ($F_{1,6} = 8.14$, $P < 0.05$) and also had higher breaking points on the progressive-ratio schedule ($F_{1,12} = 22.07$, $P < 0.0005$). In contrast, none of these behaviors increased over time in LRein rats, and in fact they tended to decrease. Finally, correlation analyses revealed that each addiction-like behavior strongly predicts the propensity to reinstatement (persistence in drug seeking, $r = 0.96$; resistance to punishment, $r = 0.67$; motivation for the drug, $r = 0.79$; $P < 0.001$ in all cases). A regression analysis including the three addiction-like behaviors as independent variables showed a multiple R equal to 0.82 ($P < 0.001$).

In a second experiment ($n = 15$), we assessed whether addiction-like behaviors could also be related to the propensity to reinstatement after a longer period of withdrawal (30 days). This time, reinstatement of responding induced by both cocaine and a cocaine-associated conditioned stimulus (CS) was studied (*15*). HRein rats ($n = 6$) showed higher levels of reinstatement responding induced by cocaine ($F_{3,30} = 4.07$, $P < 0.01$) and by the CS ($F_{1,10} = 4.62$, $P < 0.05$) than did LRein rats ($n = 6$) (Fig. 2, D and E) (*15*). Again (Fig. 2, A to C) (*15*), HRein rats displayed higher levels of addiction-like behaviors than did LRein rats (group effect for each behavior,

INSERM U588, Laboratoire de Physiopathologie des Comportements, Bordeaux Institute for Neurosciences, University Victor Segalen–Bordeaux 2, Domaine de Carreire, Rue Camille Saint-Saëns, 33077 Bordeaux Cedex, France.

*To whom correspondence should be addressed. E-mail: piazza@bordeaux.inserm.fr

$F_{1,10} = 7.09$ to 13.73, $P < 0.05$ to 0.005).

In humans, the diagnosis of addiction is performed by counting the number of diagnostic criteria that are met by an individual subject; a positive diagnosis is made when a preestablished number of criteria are found (*12*). We used a similar approach in rats by scoring them for each of the three addiction-like behaviors. For this analysis we added rats from a third experiment ($n = 26$) to increase the total number of subjects ($n = 58$) that completed the SA procedure. An individual was considered positive for an addiction-like criterion when its score for one of the three addiction-like behaviors was in the 66th to 99th percentile of the distribution (*15*). This allowed us to separate our sample of rats into four groups according to the number of positive criteria met (zero to three). The intensity of the three addiction-like behaviors was proportional to the number of criteria met by the subject (criteria effect for each behavior, $F_{3,54} = 16.99$ to 30.7, $P < 0.0001$) (Fig. 3, A to C) (*15*). Strikingly, the group that met all three positive criteria represented 17% of the entire sample (Fig. 3D), a percentage similar to that of human cocaine users diagnosed as addicts (*13*). Finally, despite this profound difference in addiction-like behavior scores, rats showing zero or three addiction-like behaviors did not differ on intake of cocaine during the entire SA period (Fig. 3E) or on sensitivity to the unconditioned effects of the drug, as measured by locomotion during SA (Fig. 3F) (*15*).

A factor analysis was then performed to determine whether the three addiction-like behaviors and the level of responding during extinction (*15*) were indices of two different underlying constructs. Extinction conditions allow for the measurement of persistence of responding for the drug when it is no longer available. Continued responding under these conditions is considered a measure of impulsivity/disinhibition (*18*), a general factor that could influence addiction-like behaviors (*19*, *20*). Remarkably, the factor analysis showed that the three addiction-like behaviors loaded equally on one factor ($r = 0.70$ to 0.88) and extinction loaded on a second independent factor ($r = 0.94$), with minimal cross-loading (Fig. 4A) (tables S1 and S2) (*15*). These findings indicate that the three addiction-like behaviors are measures of a single factor that may reflect compulsive drug use.

Finally, complementary behavioral tests were performed in rats from a fourth experiment ($n = 44$). These studies (*15*) confirmed that other dimensions previously related to vulnerability to drugs (*19–24*) do not explain individual differences in addiction-like behaviors. For example, rats with zero or three positive criteria did not differ (Fig. 4, B and C) (*15*) with respect to spontaneous motor activity (*19–22*) and anxiety-like behaviors (*23*, *24*). Similarly, a higher sensitivity to the unconditioned effects of the drug did not seem to be involved, because drug seeking persisted in a drug-free state ($F_{1,23} = 8.74$, $P < 0.005$) (Fig. 4D) (*15*). In contrast, as predicted by DSM-IV criteria, addiction-like behaviors were associated with difficulty in limiting drug intake when access to the drug was prolonged ($F_{1,23} = 4.4$, $P < 0.05$) (Fig. 4E).

These experiments show that after a prolonged period of SA, addiction-like behaviors can be found in rats. Although it is always difficult to translate findings from rats to humans, our data show striking similarities between the two species. Some rats develop behaviors similar to the diagnostic criteria for addiction described in the DSM-IV. Addiction-like behaviors are not present after a short period of SA but develop, as does addiction in humans, only after a prolonged exposure to the drug. Furthermore, as do human addicts, rats showing an addiction-like behavior have a

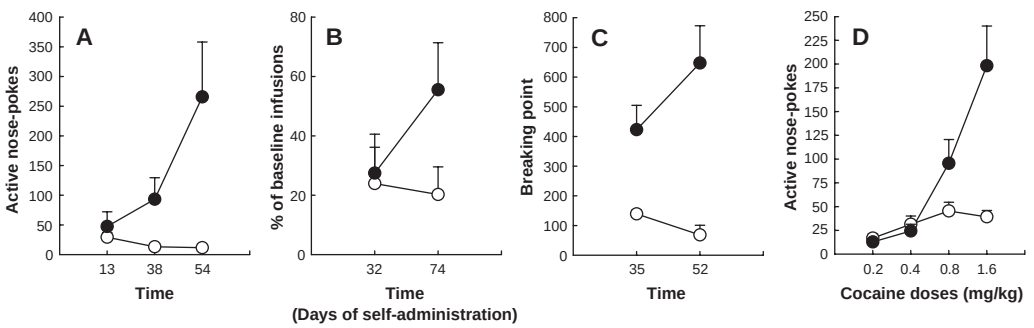


Fig. 1. Development of addiction-like behaviors over subsequent cocaine SA sessions in rats showing high (●, HRein) or low (○, LRein) cocaine-induced reinstatement after 5 days of withdrawal. (**A**) Persistence in drug seeking, as measured by number of nose-pokes in the cocaine-associated device during the no-drug period. (**B**) Resistance to punishment, as measured by change in the number of cocaine self-infusions (expressed as percentage of baseline SA) when cocaine delivery was associated with an electric shock. (**C**) Motivation for the drug, as measured by the breaking point during a progressive-ratio schedule. (**D**) Drug-induced reinstatement, as measured by number of nose-pokes in the drug-associated device as a function of the priming dose of cocaine. LRein and HRein contained the rats ($n = 7$ per group) with the lowest and highest reinstatement, respectively, induced by cocaine infusion at 1.6 mg/kg.

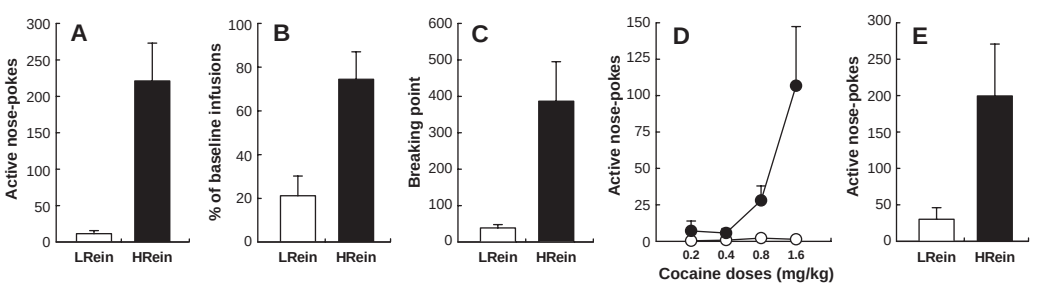


Fig. 2. Development of addiction-like behaviors over subsequent cocaine SA sessions in rats showing high (HRein) or low (LRein) cocaine-induced reinstatement after a 30-day withdrawal period. (**A**) Persistence in drug seeking, as measured by number of nose-pokes in the cocaine-associated device during the no-drug period of the 54th SA session. (**B**) Resistance to punishment, as measured by change in the number of cocaine self-infusions (expressed as percentage of baseline SA) when cocaine delivery was associated with an electric shock during the 72nd SA session. (**C**) Motivation for the drug, as measured by the breaking point during the progressive-ratio schedule conducted during the 60th SA session. (**D**) Drug-induced reinstatement, as measured by number of nose-pokes in the drug-associated device as a function of the priming dose of cocaine. (**E**) Reinstatement induced by a conditioned stimulus (CS), as measured by the number of nose-pokes in the drug-associated device when responding was associated with the contingent presentation of the CS. LRein and HRein contained the rats ($n = 6$ per group) with the lowest and highest reinstatement, respectively, induced by cocaine infusion at 1.6 mg/kg. Tests for cocaine- and CS-induced reinstatements were performed after 30 and 32 days of withdrawal, respectively, using a latin square design.

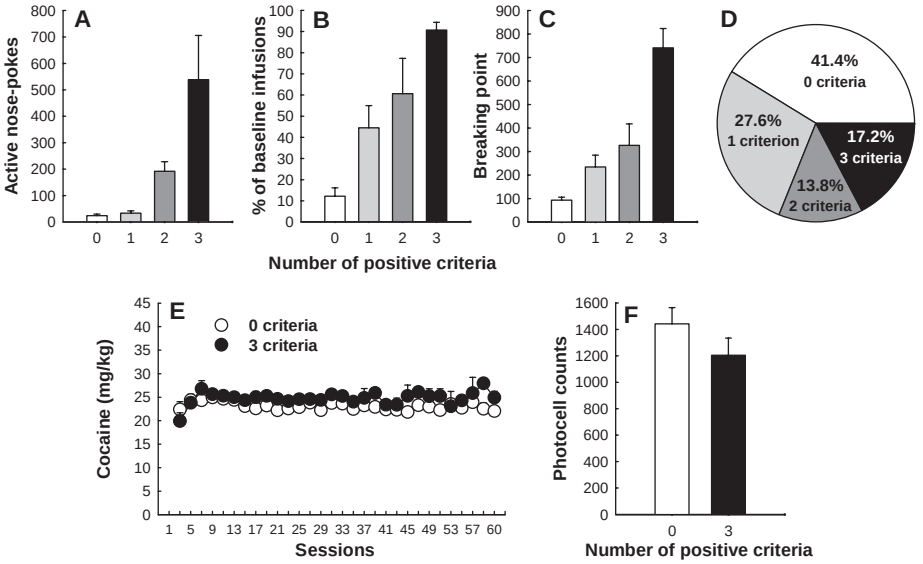


Fig. 3. (**A** to **D**) Addiction-like behaviors in rats positive for the presence of zero, one, two, or three addiction-like criteria. An individual was considered positive for an addiction-like criterion when its score for one of the three addiction-like behaviors was in the 66th to 99th percentile of the distribution. (A) Persistence in drug seeking, as measured by number of nose-pokes in the cocaine-associated device during the no-drug period of the 54th SA session. (B) Resistance to punishment, as measured by change in the number of cocaine self-infusions (expressed as percentage of baseline SA) when cocaine delivery was associated with an electric shock between the 72nd and 74th SA sessions. (C) Motivation for the drug, as measured by the breaking point during a progressive-ratio schedule performed between the 52nd and 60th SA sessions. (D) Percentage of the total population (n = 58) of rats positive for zero, one, two, or three addiction-like criteria. (**E** and **F**) Drug intake and motor activity during baseline SA in rats positive for the presence of zero or three addiction-like criteria. (E) Cocaine intake per session during baseline SA sessions (every other session is represented). (F) Horizontal motor activity during SA, as measured by number of photocell beam breaks. Results are expressed as the mean over three baseline SA sessions (between sessions 49 and 59).

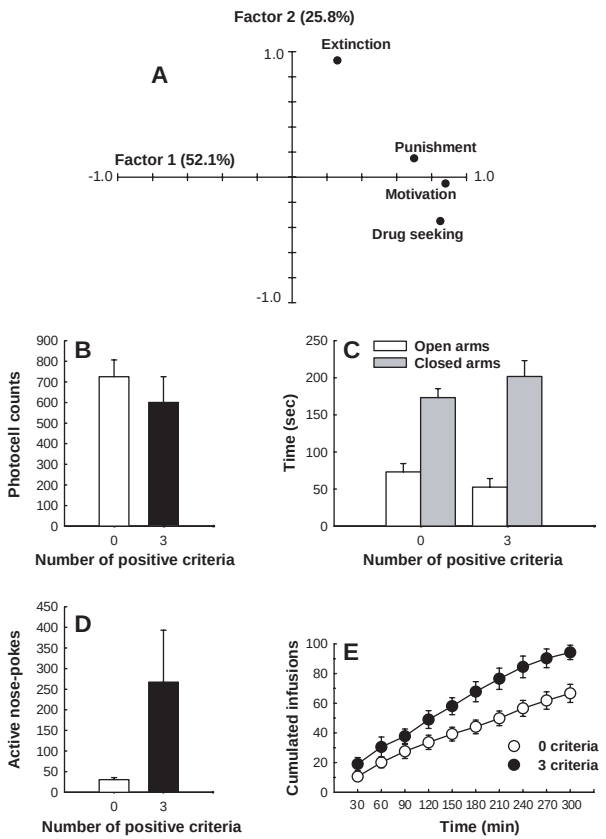


Fig. 4. (**A**) Factor analysis of SA variables. Two factors were extracted; factor 1 is represented by the horizontal axis and factor 2 by the vertical axis. Factor 1 (compulsive drug intake) accounts for 52.1% of the total variance, factor 2 (extinction) for 25.8%. The locations of the variables (●) correspond to the following parameters: Persistence in drug seeking was measured by the number of nose-pokes in the cocaine-associated device during the no-drug period of the 54th SA session. Resistance to punishment was measured by change in the number of cocaine self-infusions (expressed as percentage of baseline SA) when cocaine delivery was associated with an electric shock between the 72nd and 74th SA sessions. Motivation for the drug was measured by the breaking point during a progressive-ratio schedule performed between the 52nd and 60th SA sessions. Extinction was measured by the number of active nose-pokes during a 1-hour extinction session conducted between the 60th and 63rd sessions. (**B** and **C**) Measures of other potentially drug-related behaviors in rats positive for the presence of zero or three addiction-like criteria. (B) Spontaneous horizontal motor activity, as measured by number of photocell beam breaks exhibited during a 2-hour exposure to a novel environment. (C) Anxiety-related behavior, as measured by the comparison of the time spent in the open versus the closed arms during a 5-min exposure to an elevated plus-maze. (**D** and **E**) Measure of drug seeking in a drug-free state and of drug taking during extended access to cocaine in rats positive for the presence of zero or three addiction-like criteria. (D) Persistence in drug seeking in a drug-free state, as measured by the number of nose-pokes in the cocaine-associated device when a no-drug period precedes the SA session (mean of five consecutive tests performed between the 47th and 58th SA sessions). (E) SA during extended access to the drug. Cocaine was continuously accessible for 5 hours, and SA was estimated by the cumulated number of self-infusions over time.

high propensity to relapse even after a long period of withdrawal. Finally, the percentage of rats (17%) that show a high score for all three addiction-like criteria is similar to the percentage (15%) of human cocaine users diagnosed as addicts (*13*).

It could seem surprising that the capacity of drugs of abuse to induce addiction-like behavior exists across species. As already mentioned, however, voluntary intake of drugs abused by humans is present in several species (*1–6*). Drugs of abuse have reinforcing effects by activating endogenous reward systems that are similar in different species. Thus, the mechanisms mediating the neuroadaptations induced by chronic drug exposure and their behavioral consequences (addiction) may be similar in different species. Indeed, preliminary results show similar changes in brain activity between rats showing addiction-like behaviors and human addicts (*15*).

Our results allow us to propose a unified vision of the origin of addiction that integrates the experimental and clinical perspectives. The major hypotheses driving experimental research consider the degree of drug exposure as the key factor leading to addiction (*7–10*, *25*). By contrast, clinical visions of drug abuse have been progressively shifting weight from the role of drug exposure to the role of the higher vulnerability to drugs in certain individuals (*26–30*). Our data indicate that addiction results from the interaction of these two variables: (i) the degree of exposure to drugs (because addiction-like behavior appears only after extended access to cocaine), and (ii) the degree of vulnerability in the exposed individual (because, despite a similar drug intake in all subjects, addiction-

like behavior appears only in a few). It is thus the interaction between a long exposure to drug and a vulnerable phenotype, not one or the other factor in itself, that seems to determine the development of addiction.

References and Notes

1. T. Kusayama, S. Watanabe, *Neuroreport* **11**, 2511 (2000).
2. C. I. Abramson *et al*., *Alcohol. Clin. Exp. Res.* **24**, 1153 (2000).
3. C. R. Schuster, T. Thompson, *Annu. Rev. Pharmacol.* **9**, 483 (1969).
4. R. Pickens, W. C. Harris, *Psychopharmacologia* **12**, 158 (1968).
5. J. R. Weeks, *Science* **138**, 143 (1962).
6. S. R. Goldberg, J. H. Woods, C. R. Schuster, *Science* **166**, 1306 (1969).
7. E. J. Nestler, G. K. Aghajanian, *Science* **278**, 58 (1997).
8. S. E. Hyman, R. C. Malenka, *Nature Rev. Neurosci.* **2**, 695 (2001).
9. G. F. Koob, M. Le Moal, *Science* **278**, 52 (1997).
10. B. J. Everitt, M. E. Wolf, *J. Neurosci.* **22**, 3312 (2002).
11. R. A. Wise, *Neuron* **36**, 229 (2002).
12. *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, Washington, DC, ed. 4, revised version, 2000).
13. J. C. Anthony *et al*., *Exp. Clin. Psychopharmacol.* **2**, 244 (1994).
14. W. DeJong, *Int. J. Addict.* **29**, 681 (1994).
15. See supporting data at *Science* Online.
16. N. R. Richardson, D. C. Roberts, *J. Neurosci. Methods* **66**, 1 (1996).
17. Y. Shaham, U. Shalev, L. Lu, H. De Wit, J. Stewart, *Psychopharmacology* **168**, 3 (2003).
18. Y. Shaham, S. Erb, J. Stewart, *Brain Res. Rev.* **33**, 13 (2000).
19. R. N. Cardinal, D. R. Pennicott, C. L. Sugathapala, T. W. Robbins, B. J. Everitt, *Science* **292**, 2499 (2001).
20. R. Ito, T. W. Robbins, B. J. Everitt, *Nature Neurosci.* **7**, 389 (2004).
21. P. V. Piazza, J.-M. Deminière, M. Le Moal, H. Simon, *Science* **245**, 1511 (1989).
22. P. V. Piazza, V. Deroche-Gamonet, F. Rougé-Pont, M. Le Moal, *J. Neurosci.* **20**, 4226 (2000).
23. J. R. Homberg *et al*., *Eur. J. Neurosci.* **15**, 1542 (2002).
24. R. Spanagel *et al*., *Psychopharmacology* **122**, 369 (1995).
25. T. E. Robinson, K. C. Berridge, *Brain Res. Rev.* **18**, 247 (1993).
26. C. P. O'Brien, R. N. Ehrman, J. N. Terns, in *Behavioral Analysis of Drug Dependence*, S. R. Goldberg, I. P. Stolerman, Eds. (Academic Press, New York, 1986), pp. 329–356.
27. H. de Wit, E. H. Uhlenhuth, C. E. Johanson, *Drug Alcohol Depend.* **16**, 341 (1986).
28. M. A. Enoch, *Am. J. Pharmacogenomics* **3**, 217 (2003).
29. T. J. Crowley *et al*., *Drug Alcohol Depend.* **49**, 225 (1998).
30. D. M. Ferguson, L. J. Horwood, M. T. Lynskey, P. A. Madden, *Arch. Gen. Psychiatry* **60**, 1033 (2003).

31. We thank E. Balado for precious technical help and D. H. Epstein for insightful comments of this manuscript. Supported by INSERM, Bordeaux Institute for Neurosciences (IFR8), University Victor Segalen–Bordeaux 2, and Région Aquitaine.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/1014/DC1
Materials and Methods
SOM Text
Tables S1 and S2

12 April 2004; accepted 7 July 2004

Drug Seeking Becomes Compulsive After Prolonged Cocaine Self-Administration

Louk J. M. J. Vanderschuren*† and Barry J. Everitt

Compulsive drug use in the face of adverse consequences is a hallmark feature of addiction, yet there is little preclinical evidence demonstrating the actual progression from casual to compulsive drug use. Presentation of an aversive conditioned stimulus suppressed drug seeking in rats with limited cocaine self-administration experience, but no longer did so after an extended cocaine-taking history. In contrast, after equivalent extended sucrose experience, sucrose seeking was still suppressed by an aversive conditioned stimulus. Persistent cocaine seeking in the presence of signals of environmental adversity after a prolonged cocaine-taking history was not due to impaired fear conditioning, nor to an increase in the incentive value of cocaine, and may reflect the establishment of compulsive behavior.

Compulsive drug seeking and drug taking distinguishes drug addicts from casual drug users. Addicts display drug-dominated, inflexible behavior and are unable to shift their thoughts and behavior away from drugs and drug-related activities. Even with awareness of the deleterious consequences of this drug-centered behavior, addicts have enormous difficulty in abstaining from drug seeking and use (*1*, *2*). Several hypotheses try to explain the occurrence of compulsive drug use; it may reflect the establishment of an automatic stimulus-response habit (*3*, *4*), drug-induced loss of impulse control (*5*), sensitization of an incentive ("wanting") system (*6*), or disruption of hedonic homeostasis (*7*). Remarkably, there is little evidence from animal studies demonstrating the actual progression from casual to compulsive drug use, although drug intake in rats escalates after weeks of prolonged drug self-administration (*8*). Modeling compulsive drug seeking in animals would clarify our understanding of the neuropsychology of drug addiction and may also lead to the development of novel treatments.

Here, we tested the hypothesis that an extended drug-taking history renders drug seeking impervious to environmental adversity (such as signals of punishment), capturing one element of its compulsive nature (*1*). Appetitive behavior for natural and drug rewards is readily suppressed by aversive environmental stimuli or outcomes, a phenomenon termed conditioned suppression (*9*–*11*). We investigated whether the ability of a footshock-paired conditioned stimulus (CS) to suppress cocaine-seeking behavior diminishes after a prolonged cocaine-taking history and whether this reduced susceptibility to conditioned suppression also followed a similarly prolonged history of seeking sucrose, a high-incentive natural reinforcer.

In experiment 1, 21 rats were trained to self-administer cocaine under a heterogeneous seeking-taking chain schedule (*12*), in which drug seeking and taking are separate acts (*13*). Thus, meeting a response requirement on one lever (the seeking lever) in an operant chamber never resulted in drug, but instead gave access to a second lever (the taking lever), responding on which resulted in an intravenous infusion of cocaine. Immediately after the rats reached training criterion on this schedule—that is, after a limited cocaine-taking history—12 rats received tone-footshock pairings (the CS-shock group), whereas the other nine rats received presentations of the same tone not paired with footshock (the control group). Several days later, conditioned suppression of drug seeking was assessed in a session in which the rats had access to the seeking lever only and the footshock CS was presented for three 2-min periods interspersed with three 2-min periods when no CS was presented (*13*). The CS-shock group showed a profound conditioned suppression of drug seeking during presentation of the CS [$F(\mathrm{CS}) = 5.32$, $P < 0.05$; $F(\mathrm{CS} \times \mathrm{group}) = 25.65$, $P < 0.001$] (Fig. 1A). There was also a marked increase in the time taken to make the first seeking response (seeking latency) in the CS-shock group [$F(\mathrm{group}) = 7.43$, $P < 0.01$] (Fig. 1B). Thus, in rats with limited cocaine self-administration experience, drug seeking was greatly suppressed by presentation of an aversive CS, showing that it was sensitive to an adverse outcome.

Next, we tested whether cocaine seeking would become less susceptible to an aversive CS after prolonged cocaine self-administration

Department of Experimental Psychology, University of Cambridge, Cambridge CB2 3EB, UK.

*Present address: Rudolf Magnus Institute of Neuroscience, Department of Pharmacology and Anatomy, University Medical Center Utrecht, 3584 CG Utrecht, Netherlands.

†To whom correspondence should be addressed. E-mail: l.j.m.j.vanderschuren@med.uu.nl

experience. The same 21 rats were therefore allowed a further 20 cocaine self-administration sessions. These included eight "extended-access" sessions in which the rats could respond for cocaine under a simple continuous reinforcement (FR1) schedule for a maximum of 80 infusions, thereby greatly increasing the extent of cocaine-taking experience and cocaine exposure. Subsequently, the rats were reconditioned (CS-shock pairings) and tested under circumstances identical to those in the previous phase of the experiment. Rats with an extended cocaine-taking history showed virtually no conditioned suppression of drug seeking [F(CS) = 0.47, not significant (n.s.); F(CS × group) = 1.55, n.s.] (Fig. 1C), although the response latency in the CS-shock group was still somewhat increased [F(group) = 8.39, $P < 0.01$] (Fig. 1D).

An advantageous characteristic of the seeking-taking chain schedule used is that the rate of responding on the seeking lever is a function of reinforcer magnitude. Thus, rats responding for higher unit doses of cocaine or higher concentrations of sucrose show increased rates of responding on the seeking lever (*12*); seeking rate can therefore be used as a measure of the incentive value of the reinforcer. A plausible explanation for the reduced susceptibility of cocaine seeking to presentation of the footshock CS is that the incentive value of cocaine increases after prolonged cocaine exposure. Therefore, rats may have been less prepared to reduce their cocaine-seeking rates when faced with signals of an adverse environmental event because cocaine had become a more valuable commodity. To test this possibility, we compared cocaine-seeking rates after limited and extended cocaine exposure but before CS-shock conditioning. Remarkably, the seeking rates were not different under these conditions; that is, they were not affected by the amount and duration of cocaine exposure [F(experiment phase) = 0.01, n.s.] (Fig. 1E), suggesting that the incentive value of cocaine had not changed over the course of the extended self-administration history.

In experiment 2, we aimed to exclude the possibility that different degrees of between-session extinction of the footshock CS ac-

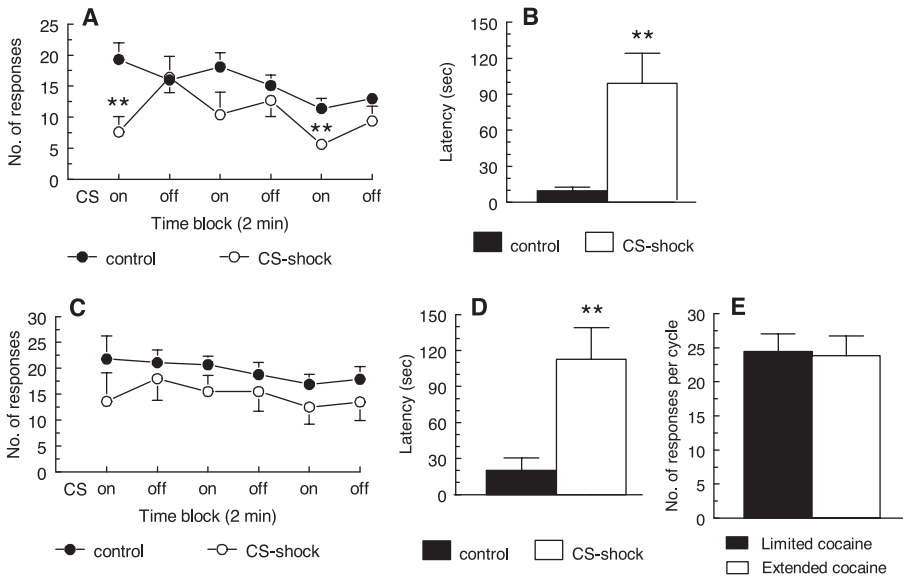


Fig. 1. Presentation of an aversive CS suppresses cocaine seeking after limited (**A** and **B**) but not prolonged (**C** and **D**) cocaine self-administration, independent of changes in the incentive value of cocaine (**E**). (A) Mean (± SEM) cocaine-seeking responses per 2-min interval in the CS-shock and control groups after limited cocaine exposure, with the aversive CS on or off during alternating 2-min periods. $^{**}P < 0.01$ (Student-Newman-Keuls). (B) Latency to make the first seeking response during the test for conditioned suppression of cocaine seeking in the CS-shock and control groups after limited cocaine exposure. $^{**}P < 0.01$ [analysis of variance (ANOVA)]. (C) Mean (± SEM) cocaine-seeking responses per 2-min interval in the CS-shock and control groups after extended cocaine exposure, with the aversive CS on or off during alternating 2-min periods. (D) Seeking latency during the test for conditioned suppression of cocaine seeking in the CS-shock and control groups after extended cocaine exposure. $^{**}P < 0.01$ (ANOVA). (E) Mean (± SEM) number of cocaine-seeking responses per seeking-taking chain cycle after limited and extended cocaine self-administration.

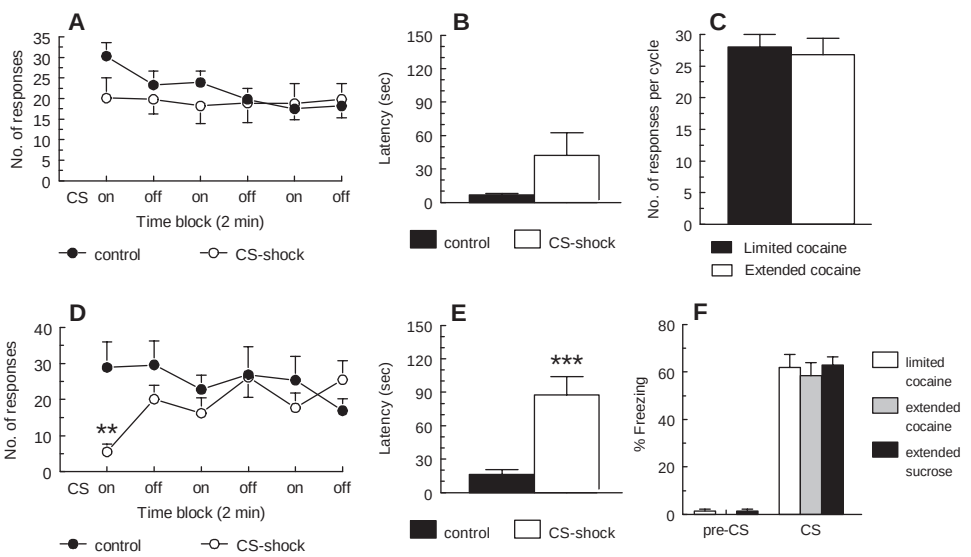


Fig. 2. Presentation of an aversive CS does not suppress cocaine seeking after prolonged cocaine self-administration (**A** and **B**), independent of changes in the incentive value of cocaine (**C**). Presentation of an aversive CS suppresses sucrose seeking after prolonged sucrose self-administration (**D** and **E**). The differences in conditioned suppression were not the result of differences in conditioned fear (**F**). (A) Mean (± SEM) cocaine-seeking responses per 2-min interval in the CS-shock and control groups after extended cocaine exposure, with the aversive CS on or off during alternating 2-min periods. (B) Seeking latency during the test for conditioned suppression of cocaine seeking in the CS-shock and control groups after extended cocaine exposure. (C) Mean (± SEM) number of cocaine-seeking responses per seeking-taking chain cycle after limited and extended cocaine self-administration. (D) Mean (± SEM) sucrose-seeking responses per 2-min interval in the CS-shock and control groups after extended sucrose exposure, with the aversive CS on or off during alternating 2-min periods. $^{**}P < 0.01$ (Student-Newman-Keuls). (E) Seeking latency during the test for conditioned suppression of sucrose seeking in the CS-shock and control groups after extended sucrose exposure. $^{***}P < 0.001$ (ANOVA). (F) Conditioned freezing to a footshock CS in rats with limited cocaine, extended cocaine, and extended sucrose self-administration experience. Percentage of time spent freezing was scored for 2 min before (pre-CS, left panel) and during 2 min of presentation of the CS (right panel).

counted for the diminished conditioned suppression seen after extended, as compared to limited, cocaine exposure—that is, that rats had learned during the suppression tests, when the CS was repeatedly presented, that it no longer predicted footshock when subsequently presented in the self-administration environment. Therefore, rats (CS-shock group, $n = 11$; control group, $n = 12$) were trained to self-administer cocaine under conditions identical to those in experiment 1. However, the first suppression test was omitted, so that rats were conditioned and tested only after extended cocaine exposure, including eight extended-access sessions (*13*). Unlike the limited-exposure rats, and identical to the extended-exposure rats in experiment 1, cocaine seeking in the CS-shock group was not suppressed at all during presentation of the footshock CS [F(CS) = 2.53, n.s.; F(CS × group) = 4.40, $P < 0.05$] (Fig. 2A). Moreover, the seeking latency in the CS-shock group was no different from that in the control group [F(group) = 2.15, n.s.] (Fig. 2B). Consistent with experiment 1, there was also no change in the incentive value of cocaine, as assessed by the rates of seeking before and after the eight extended-access sessions [F(experiment phase) = 0.08, n.s.] (Fig. 2C). Thus, cocaine seeking is greatly suppressed by a footshock CS, but only in rats with a limited cocaine-taking history. This indicates that the flexibility of drug seeking strongly depends on the extent of drug self-administration experience.

Conditioned suppression of appetitive behavior has been readily observed in a variety of settings and is commonly used as an index of conditioned fear in animals responding for natural reinforcers, such as food (*9*, *10*). However, it is not known whether food seeking can also become insensitive to the suppressive effects of an aversive CS after prolonged experience. In experiment 3, we therefore replicated experiment 2 with rats trained to seek and ingest sucrose (*13*). Rats (CS-shock group, $n = 11$; control group, $n = 11$) were trained to respond for a sucrose solution under the same seeking-taking chain schedule. To make an optimal comparison between cocaine and sucrose self-administration, we chose a unit amount and concentration of sucrose that led to rates of responding on the seeking lever comparable to those in experiment 2 (experiment 3 versus experiment 2: sucrose seeking, 13.9 ± 1.8 responses/min; cocaine, 12.6 ± 1.7 responses/min). In addition, the sucrose-trained rats were trained for a comparable number of sessions and received a comparable number of total reinforcer presentations before assessing conditioned suppression (total number of sucrose reinforcers in experiment 3: 1148 ± 29; total number of cocaine reinforcers in experiment 2: 1110 ± 14). After extended experience of sucrose seeking, profound conditioned suppression during presentation of the footshock CS was still observed [F(CS) = 5.31, $P < 0.05$; F(CS × group) = 8.35, $P < 0.01$] (Fig. 2D), together with a marked increase in the seeking latency in the CS-shock group [F(group) = 17.27, $P < 0.001$] (Fig. 2E). Thus, lengthy training under this seeking-taking chain schedule does not itself result in diminished sensitivity of appetitive behavior to presentation of an aversive CS. Rather, these data suggest that the nature of the reinforcer (drug versus natural) determines whether compulsive behavior resistant to adverse environmental events will develop after similarly prolonged periods of self-administration (*14*, *15*).

It is important to exclude the possibility that the failure to observe conditioned suppression in the extended cocaine exposure group reflected weaker fear conditioning. Therefore, the long-term sucrose-trained rats from experiment 3, the long-term cocaine-trained rats from experiment 2, and a new group of rats with limited cocaine exposure (as in experiment 1) all underwent fear conditioning and were tested for conditioned freezing to a discrete auditory (clicker) CS (*9*, *13*, *16*). Twenty-four hours after conditioning, the rats were placed in the training context and after 2 min, the clicker CS was played for 2 min (*13*). Rats in all three groups exhibited profound freezing during the CS [F(CS) = 552.2, $P < 0.01$], and there were no differences in fear behavior among the three groups during the pre-CS or CS periods [F(CS × group) = 0.77, n.s.; pre-CS freezing: F(group) = 0.003, n.s.; CS freezing: F(group) = 0.82, n.s.] (Fig. 2F). Thus, the differences in conditioned suppression cannot be attributed to altered pain sensitivity or an inability to encode or express a CS-footshock association after extended cocaine exposure. Conditioned suppression and conditioned freezing are well known to be highly correlated (*9*), but this correlation between freezing and the suppression of cocaine-seeking behavior was lost in rats with a prolonged cocaine-taking history because they were still fearful, yet their appetitive behavior was not affected by an aversive CS during the unflagging pursuit of drug.

Dysfunction of prefrontal cortical-striatal systems is likely to underlie loss of control over drug use. These systems subserve the coordination of goal-directed and habitual appetitive behavior (*17*, *18*) and have been implicated in both obsessive-compulsive disorder (*19*) and drug addiction (*20*). Indeed, animal studies suggest a critical role for the prefrontal cortex in drug seeking (*21*). Moreover, functional neuroimaging studies in human drug addicts have consistently shown activation of the orbitofrontal and dorsolateral prefrontal cortex during cocaine craving (*20*), and cocaine addicts are impaired in cognitive and decision-making abilities that depend on the orbital and other prefrontal cortical areas (*22*).

Our results show that cocaine seeking can be suppressed by presentation of an aversive CS, but after extended exposure to self-administered cocaine, drug seeking becomes impervious to adversity. Interestingly, inflexible drug seeking appears to develop with prolonged drug-taking experience independently of alterations in the incentive value of cocaine. The attenuated conditioned suppression of seeking does not occur after identically prolonged exposure to sucrose, which suggests that appetitive behavior may more readily become resistant to aversive environmental events when directed toward obtaining drugs rather than natural reinforcers. We therefore conclude that a prolonged cocaine self-administration history endows drug seeking with an inflexible, compulsive dimension.

References and Notes

1. *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, Washington, DC, ed. 4, 1994).
2. C. P. O'Brien, A. T. McLellan, *Lancet* **347**, 237 (1996).
3. S. T. Tiffany, *Psychol. Rev.* **97**, 147 (1990).
4. B. J. Everitt, A. Dickinson, T. W. Robbins, *Brain Res. Rev.* **36**, 129 (2001).
5. J. D. Jentsch, J. R. Taylor, *Psychopharmacology* **146**, 373 (1999).
6. T. E. Robinson, K. C. Berridge, *Brain Res. Rev.* **18**, 247 (1993).
7. G. F. Koob, M. Le Moal, *Science* **278**, 52 (1997).
8. S. H. Ahmed, G. F. Koob, *Science* **282**, 298 (1998).
9. M. E. Bouton, R. C. Bolles, *Anim. Learn. Behav.* **8**, 429 (1980).
10. S. Killcross, T. W. Robbins, B. J. Everitt, *Nature* **388**, 377 (1997).
11. D. N. Kearns, S. J. Weiss, L. V. Panlilio, *Drug Alcohol Depend.* **65**, 253 (2002).
12. M. C. Olmstead *et al*., *Psychopharmacology* **152**, 123 (2000).
13. See supporting data at *Science* Online.
14. J. D. Berke, S. E. Hyman, *Neuron* **25**, 515 (2000).
15. E. J. Nestler, *Nature Rev. Neurosci.* **2**, 119 (2001).
16. J. E. LeDoux, A. Sakaguchi, D. J. Reis, *J. Neurosci.* **4**, 683 (1984).
17. S. Killcross, E. Coutureau, *Cereb. Cortex* **13**, 400 (2003).
18. H. H. Yin, B. J. Knowlton, B. W. Balleine, *Eur. J. Neurosci.* **19**, 181 (2004).
19. A. M. Graybiel, S. L. Rauch, *Neuron* **28**, 343 (2000).
20. R. Z. Goldstein, N. D. Volkow, *Am. J. Psychiatry* **159**, 1642 (2002).
21. P. W. Kalivas, K. McFarland, *Psychopharmacology* **168**, 44 (2003).
22. R. D. Rogers, T. W. Robbins, *Curr. Opin. Neurobiol.* **11**, 250 (2001).

23. This research was funded by an MRC Programme Grant and was conducted within the Cambridge MRC Centre for Clinical and Behavioral Neuroscience. L.J.M.J.V. was a visiting scientist from the Research Institute Neurosciences VU, Department of Medical Pharmacology, VU Medical Center, Amsterdam, supported by a Wellcome Trust Travelling Research Fellowship. We thank P. Di Ciano and J. L. C. Lee for practical assistance and help with the design of the experiments and R. N. Cardinal for additional software programming.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/1017/DC1
Materials and Methods
References

9 April 2004; accepted 2 July 2004

Visual Pattern Recognition in *Drosophila* Is Invariant for Retinal Position

Shiming Tang,[1]*† Reinhard Wolf,[2]* Shuping Xu,[1] Martin Heisenberg[2]†

Vision relies on constancy mechanisms. Yet, these are little understood, because they are difficult to investigate in freely moving organisms. One such mechanism, translation invariance, enables organisms to recognize visual patterns independent of the region of their visual field where they had originally seen them. Tethered flies (*Drosophila melanogaster*) in a flight simulator can recognize visual patterns. Because their eyes are fixed in space and patterns can be displayed in defined parts of their visual field, they can be tested for translation invariance. Here, we show that flies recognize patterns at retinal positions where the patterns had not been presented before.

In the flight simulator (Fig. 1A), the fly's (*Drosophila melanogaster*) head and thorax and, hence, its eyes are fixed in space while its yaw torque can still control the angular velocity of a panorama surrounding it (*1*). If the panorama displays different patterns the fly can be trained to discriminate them (Fig. 1B) (*2*). The flight simulator lends itself to an investigation of translation invariance as patterns rotate around the fly at a fixed height and can be vertically displaced between training and test. To our surprise, flies failed in such tests to recognize patterns shifted up or down by 9° or more after training. Pattern recognition seemed to require the same retinal coordinates for acquisition and retrieval (*3*–*5*). This finding was in line with earlier experiments in ants, which had failed to show interocular transfer for landmark recognition (*6*).

Subsequent studies (*7*–*9*) identified some of the pattern parameters (features) the flies used for discrimination. These were size, color, vertical compactness, and vertical position of the centers of gravity (COGs) of the patterns in the panorama. For many pattern pairs carrying none of these features, no conditioned discrimination could be detected, although flies often discriminated them spontaneously (*7*). In the earlier quest for translation invariance (*3*–*5*), flies had been conditioned to discriminate patterns solely by the vertical position of their COGs. For instance, if in Fig. 1A the vertical positions of the COGs of upright and inverted Ts were aligned, flies were unable to discriminate them after conditioning [shown for triangles in ref. (*7*) and discussed in Supplement]. This raised the possibility that perhaps vertical displacement specifically interfered with the feature "vertical position" (*10*, *11*).

Only two of the four parameters, size and color, are independent of the vertical position of the pattern elements in the arena. These, therefore, were chosen in the present study. To test for conditioned discrimination of (horizontal) size (Fig. 2A, left dotted bars), we presented two black rectangles of the same height but differing by about a factor of two in width in neighboring quadrants. They were all shown at the same vertical position in the arena slightly above the fly's horizon. Flies readily learned to avoid the larger or smaller figure after the training ($P < 0.001$). Next, these patterns were vertically displaced between training and test (*12*). In contrast to the earlier experiments with patterns differing in the vertical position of their COGs, no decrement of the memory score was observed after a vertical displacement of $\Delta H = 20°$ (Fig. 2A, right cross-hatched bars). In the same way, we tested color. Flies remembered blue and green rectangles of the same size and presented at the same vertical position (*8*, *9*). They had no difficulty recognizing them after vertical displacement at the new position (Fig. 2B).

Edge orientation is a feature that has been extensively documented in the honeybee (*13*–*17*) to serve in conditioned pattern discrimination. We found a robust conditioned preference for bars tilted +45° and −45° to the vertical [Fig. 2C, left dotted bars; (*18*); but see (*7*)]. Vertical displacement of the bars after training had no significant effect on the memory score [Fig. 2C, right cross-hatched bars; (*19*)].

Next, more complex patterns were tested. The rectangles in the four quadrants in Fig. 2D were each composed of a blue and a green horizontal bar. They differed only in whether green was above blue or blue above green. In principle, flies had two options to discriminate the two figures. They could combine the two features vertical position and color to give a new feature with relational cues (e.g., "green above blue"). Alternatively, they could evaluate the two colors separately and remember for each whether the high or low rectangles were safe or dangerous. This would have different consequences in the transfer experiment. If the colors were processed separately, flies would have to rely on vertical positions and could recognize neither the green nor the blue patterns at the new retinal positions. For composite figures, however, the vertical positions would be transformed into relational cues, which might still be recognized after vertical displacement. The latter was observed (Fig. 2D, right cross-hatched bars). Apparently, within each rectangle, flies evaluated the positions of the colored pattern elements

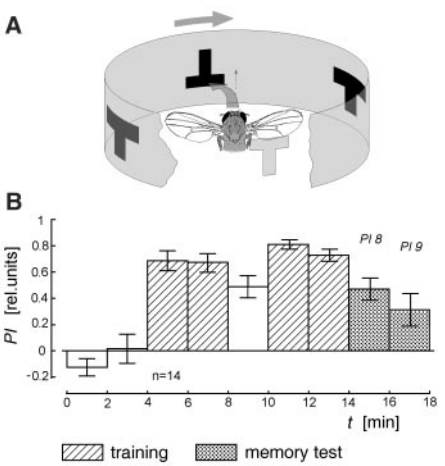


Fig. 1. Visual pattern discrimination learning in the flight simulator. Angular velocity of an artificial visual panorama is made negatively proportional to the fly's yaw torque [gray arrows in (**A**)], allowing the fly to change its flight direction (yaw torque ≠ 0), or to maintain stable orientation (yaw torque = 0) with respect to visual landmarks in the panorama (upright and upside-down T-shaped black patterns). During training, a heat beam (not shown in figure), directed to the fly's thorax and head from behind, is switched on or off at the boundaries between quadrants containing the one or the other pattern type in their center. (**B**) Standard learning experiment. Performance index (*PI*) is calculated as $PI = (t_c - t_h)/(t_c + t_h)$, where t_c is the fraction of time with heat off and t_h the remaining time with heat on in a 2-min interval. The arena is rotated to a random angular position at the beginning of each 2-min interval. Empty bars indicate test intervals without any heat; hatch bars denote training intervals. *PI* 8 and *PI* 9 (dotted bars) quantify the flies' conclusive pattern memory. Error bars are SEMs.

[1]Institute of Biophysics Academia Sinica, 15 Datun Road, Chaoyang, Beijing 100101, P.R. China. [2]Lehrstuhl für Genetik und Neurobiologie, Universität Würzburg, Biozentrum (Am Hubland), 97074 Würzburg, Germany.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: tang-shm@sohu.com, heisenberg@biozentrum.uni-wuerzburg.de

relative to each other and directed their flight with respect to these cues.

Similar relational cues were effective in Fig. 2E. Each of the alternative figures consisted of two orthogonally oriented oblique bars, one above the other. In the two figures the two orientations were exchanged. In one, the top bar was tilted by +45°, in the other by –45°. If the flies relied on the integral of the orientations of all edges in each composite figure they would not be able to discriminate the two. Flies were able to evaluate the spatial relations in the composite figures. Even for these complex patterns, translation invariance for vertical displacement was found. When the two patterns were rotated by 90°, which placed the two bars side-by-side at the same vertical position, no conditioned discrimination was obtained (Fig. 2F).

The only patterns *Drosophila* failed to recognize after vertical displacement were those that it can discriminate only by their vertical position. This suggests a special interference between the feature vertical position and the vertical displacement. Horizontal displacement of these patterns cannot be tested in the flight simulator, because horizontal motion is controlled by the fly, which has to choose a certain azimuth relative to the landmarks for its direction of flight. We therefore developed an alternative paradigm to investigate visual pattern recognition and, in particular, horizontal translation invariance. Flies were conditioned at the torque meter by heat to restrict their yaw torque range to only left or only right turns (*2*). Two patterns were displayed at stable retinal positions during training (Fig. 3), for instance at +45° and –45° from the frontal direction. Between training and memory test the patterns were exchanged. Flies shifted their restricted yaw torque range to the other side (i.e. from left turns to right turns or vice versa) (Fig. 3A).

When patterns were shifted to a new position on the same side (from ±30° to ±80° or vice versa), the yaw torque bias stayed on the side of the yaw torque range to which it had been confined during training (Fig. 3B). Obviously, flies recognized the patterns at the new retinal positions after horizontal displacement. The T-shaped patterns used in this experiment could be discriminated only by the vertical positions of their COGs, and they were the very patterns for which no translation invariance had been found in the flight simulator after vertical displacement.

All five pattern parameters tested (size, color, edge orientation, relational cues, vertical position) showed visual pattern recognition in *Drosophila* to be translation invariant. Vertical position was the only parameter that the flies could not recognize after vertical displacement, but they did recognize this parameter after a horizontal shift.

Little is known so far about translation invariance in flies. In the present study, it has been demonstrated for horizontal displacements between +45° and –45° from the frontal direction. These positions are well outside the region of binocular overlap (*20*). Hence, the pattern information generalized for position must be made available to both brain hemispheres (interocular transfer).

Our yaw torque learning paradigm reveals intriguing properties of visual processing. First, it shows that visual motion is not a prerequisite for pattern recognition. Flies with their eyes fixed in space can recognize stationary visual objects (*21*). No motion is required for the perceptual process. Although the fly can still move the optical axes of its photoreceptors by a few degrees (*22*), this is too little to generate directional motion. Moreover, flies can recognize visual patterns in the flight simulator if during acquisition these are kept stationary (*23*). In the present experiment, the patterns were stationary even

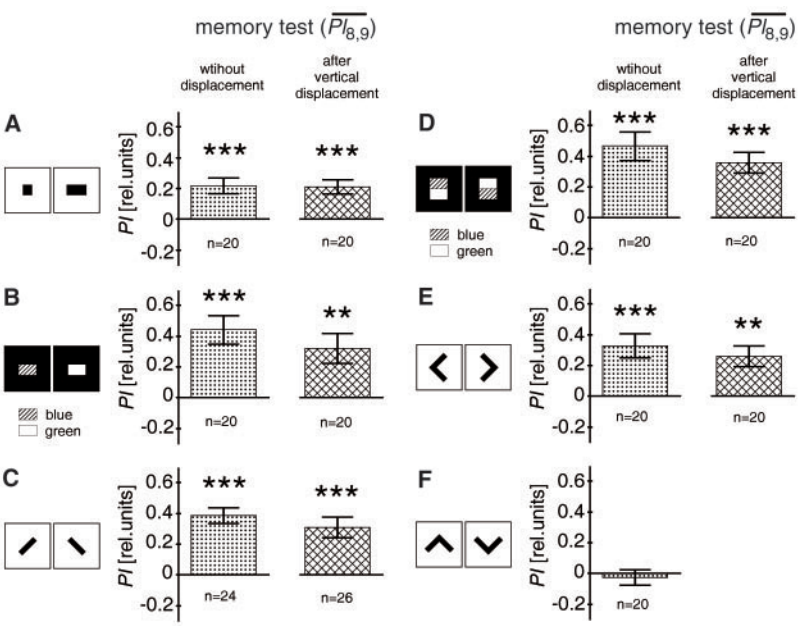


Fig. 2. Pattern discrimination learning and retinal transfer with vertical displacements for various parameters. Flies are trained with different pairs of patterns following the protocol of the standard learning experiment (see Fig.1). Bar graphs: (dotted) pattern memory tested without vertical displacement; (cross-hatched) the complete panorama was shifted upward by 20° after the last training block (at t = 14 min). Bars are averaged means of *PI* 8 and *PI* 9. Error bars are SEMs. ***$P < 0.001$; **$P < 0.01$ (From a one-sample t test, 2-tailed P value).

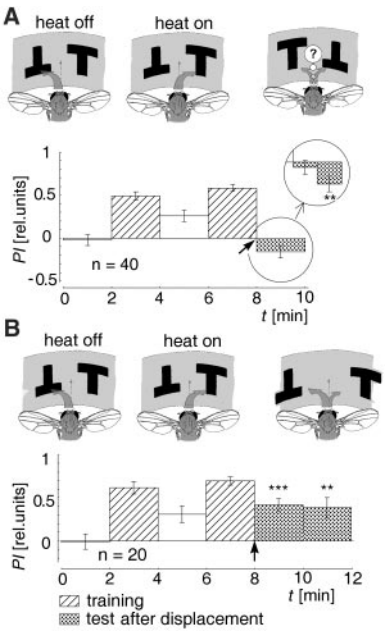


Fig. 3. Yaw torque conditioning with fixed visual patterns. Two stationary patterns are presented to the fly at +45° and –45° from the frontal direction. The fly's yaw torque is recorded and its range is divided in two domains roughly corresponding to intended left and right turns. If torque is to the right, heat is switched on, if torque is to the left, heat is switched off. Flies can learn to keep their torque persistently in the safe domain (for this experiment without visual patterns see ref. (*2*). The first test between the two training blocks is carried out with patterns in the training positions. The positive *PI* indicates that the flies continue to direct their yaw torque to the "safe" side. (**A**) The two patterns are exchanged after a second training period with the arena light switched off during the shift. Negative *PI* after the second training indicates that flies recognize the patterns in their new positions. Higher time resolution of *PI*s shows gradual transition from positive to negative *PI*s (inset in circle), which indicates different dynamics of visual and motor memory. (**B**) As a control, patterns were shifted within the same visual half-field (from $\psi = \pm 30°$ to $\psi = \pm 80°$ or vice versa). ***$P < 0.001$; **$P < 0.01$ (From a one-sample t test, two-tailed P value).

during retrieval. The fly's turning tendency indicated that it recognized the patterns.

Second, during intended turns to one side flies selectively followed the directional motion cues of landmarks on that side and neglected the symmetrical motion cues of a corresponding landmark on the other side (*24*). From the present experiment, we can deduce that the fly associated the heat with the pattern to which it tried to turn while being heated, and the no-heat condition with the pattern to which it tried to turn while flying in the cold. Because the flies were exposed to the two patterns in equivalent retinal positions, they must be able to activate a gating process for a part of the visual array in the optic lobes corresponding to one or the other of the visual half-fields. Studies of walking flies have provided similar phenomena (*25*). The ability to confine visual processing to a visual field region of choice is called selective visual attention (*26*).

Selective attention may be relevant also for the flight simulator experiment. For translation invariance in the flight simulator, the fly has to store not only a feature of a pattern for recognition ("what") but also an azimuth value for orientation ("where"). We propose that, while being heated, the fly associates the pattern that happens to be in the window of attention with the heat. In the flight simulator, the fly most of the time keeps the window of attention in a frontal position (*27*). In this way, the pattern would be labeled "dangerous if approached."

Third, the flies in the yaw torque learning paradigm not only associated heat with the turning tendency to one side but also with the pattern on the side to which they tried to turn. In the memory test after the patterns had been exchanged, the fly inverted its turning tendency. The preference for the previously "safe" torque domain quickly fades, whereas the pattern preference, expressed by the fly's yaw torque to the side of the attractive pattern, persists. If, for instance, the fly expects yaw torque to the left to entail heat but suddenly finds on that side the previously safe pattern, it overrides its negative predisposition for left turns and tries to turn into that direction. The fading of the behavioral memory component has also been reported for 3-way associations testing colors instead of patterns (*23*).

Feature detectors for edge orientation are a hallmark of mammalian visual systems (*28*) and have also extensively been studied in the honeybee (*13–17*). Like translation invariance, edge orientation is a further basic property that is shared between the visual systems of *Drosophila* and larger animals. Finally, flies also evaluate relational cues such as {A above B} versus {B above A}. So far this fascinating ability has been demonstrated only for two colors (blue and green) and two edge orientations (+45° and −45°). The negative outcome of the experiment with two horizontally arranged oblique bars in the flight simulator cannot be generalized. As mentioned above, the exclusively horizontal motion in the flight simulator may specifically interfere with this arrangement. Also horizontal compactness is not a discriminating parameter (*7*), although a grouping effect for vertical bars can be observed in fixation (*27*). In any case, the discovery of relational cues seems to vastly increase the potential number of pattern parameters the fly might be able to discriminate.

Our data suggest a basic scheme (a minimal circuit) for translation invariance. As mentioned above (*10*, *11*), the experimental paradigms conceptually demand a distinction between orientation and recognition, i.e., a where and what network (*28*, *29*). Both networks must have a centripetal (afferent) and a centrifugal (efferent) branch. The model is outlined in the supplement (fig. S1).

References and Notes

1. M. Heisenberg, R. Wolf, *J. Comp. Physiol. A Sens. Neural Behav. Physiol.* **130**, 113 (1979).
2. R. Wolf, M. Heisenberg, *J. Comp. Physiol. A Sens. Neural Behav. Physiol.* **169**, 699 (1991).
3. M. Dill, R. Wolf, M. Heisenberg, *Nature* **365**, 751 (1993).
4. M. Dill, M. Heisenberg, *Philos. Trans. R. Soc. London B Biol. Sci.* **349**, 143 (1995).
5. M. Dill, R. Wolf, M. Heisenberg, *Learn. Mem.* **2**, 152 (1995).
6. R. Wehner, M. Müller, *Nature* **315**, 228 (1985).
7. R. Ernst, M. Heisenberg, *Vision Res.* **39**, 3920 (1999).
8. S. M. Tang, A. K. Guo, *Science* **294**, 1543 (2001).
9. Note that flies discriminated the rectangles by hue rather than brightness, because varying the relative intensities of the two colors by a factor of 10 between training and test has no significant effect on memory performance.
10. M. Heisenberg, *Curr. Opin. Neurobiol.* **5**, 475 (1995).
11. In the flight simulator, an orientation task is used to measure pattern recognition and translation invariance. In order to retrieve a particular flight direction relative to the panorama in the memory test, the fly has to store the respective pattern (feature) during the training not only for recognition but also by an azimuth value for orientation (e.g., direction of flight).
12. In previous experiments (*3*, *6*), the transparency in the arena carrying the figures had been exchanged between training and test. This procedure required 30 to 60 s. Also, for half of the flies, the pattern changed from a lower to a higher position; for the other half, the sequence was the opposite. In the present experiments, the whole arena was shifted after the final training and the shift was always 20° upward. The new procedure implied that, not only the figures, but also the upper and lower margins of the arena were displaced, but the shift took only an instant and did not entail any visual disturbances from handling. Control experiments showed the same basic results with the old and new procedures. For instance, flies trained with horizontal bars at different heights (ΔH = 20°) to avoid certain flight directions were unable to retrieve this information if, after the training, the whole arena was shifted upward by 20° (*30*), this inability confirmed that the flies' pattern recognition system does not tolerate the vertical displacement if vertical position is the discriminating pattern parameter.
13. R. Wehner, *Nature* **215**, 1244 (1967).
14. R. Wehner, M. Lindauer, *Z. Vgl. Physiol.* **52**, 290 (1966).
15. J. H. van Hateren, M. V. Srinivasan, P. B. Wait, *J. Comp. Physiol. A Sens. Neural Behav. Physiol.* **167**, 649 (1990).
16. M. V. Srinivasan, S.W. Zhang, K. Witney, *Philos. Trans. R. Soc. London B Biol. Sci.***343**, 199 (1994).
17. A. Horridge, *J. Insect. Physiol.* **44**, 343 (1998).
18. M. Dill, thesis, University of Würzburg (1992).
19. As pattern recognition in honey bees is studied with freely moving animals, a formal test for translation invariance has not been possible. Nevertheless, retinotopic template-matching can be excluded as the sole mechanism (*31*). In particular, the data clearly indicate that the orientation of edges can be recognized independent of their precise location on the visual array (*32*).
20. E. Buchner, thesis, University of Tübingen (1971)
21. B. Bausenwein, R. Wolf, M. Heisenberg, *J. Neurogenet.* **3**, 87 (1985).
22. R. Hengstenberg, *Kybernetik* **9**, 56 (1971).
23. M. Heisenberg, R. Wolf, B. Brembs, *Learn. Mem.* **8**, 1 (2001).
24. R. Wolf, M. Heisenberg, *J. Comp. Physiol. A Sens. Neural Behav. Physiol.* **140**, 69 (1980).
25. S. Schuster, thesis, University of Tübingen (1996).
26. M. I. Posner, C. R. R. Snyder, B. J. Davidson, *J. Exp. Psychol. Genet.* **109**, 160 (1980).
27. M. Heisenberg, R. Wolf, *Vision in* Drosophila*: Genetics of Microbehavior*, V. Braitenberg, Ed. (Studies of Brain Function, vol. 12, Springer-Verlag, Berlin, 1984).
28. D. H. Hubel, *Eye, Brain, and Vision* (Scientific American Library, New York, 1988).
29. M. Livingstone, D. Hubel, *Science* **240**, 740 (1988).
30. S. Tang, R. Wolf, S. Xu, M. Heisenberg, unpublished results.
31. D. Efler, B. Ronacher, *Vision Res.* **40**, 3391 (2000).
32. M. V. Srinivasan, S. W. Zhang, B. Rolfe, *Nature* **362**, 539 (1993).

33. We thank B. Ronacher and L. Wiskott for valuable comments on the manuscript. This work was supported by the German Science Foundation (SFB 554) and the Multidisciplinary Research program of the Chinese Academy of Science (S.T.).

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5686/1020/DC1
Material and Methods
SOM Text
Fig. S1
References

3 May 2004; accepted 8 July 2004

New Products

BD Biosciences
For more information
800-343-2035
www.bdbiosciences.com
www.scienceproductlink.org

Peptide Hydrogel

BD PuraMatrix Peptide Hydrogel is a novel synthetic matrix used to create defined three-dimensional (3D) microenvironments for a variety of cell culture experiments. This material consists of standard amino acids and more than 99% water. Under physiological conditions, the peptide component self-assembles into a 3D hydrogel that exhibits a nanometer scale fibrous structure. In the presence of key bioactive molecules, this hydrogel promotes the attachment, growth, and differentiation of multiple cell types. It is biocompatible, resorbable, and injectable, and devoid of animal-derived material and pathogens.

Bio-Tek
For more information
888-451-5171
www.biotek.com
www.scienceproductlink.org

Ultra-Sensitive Luminometer

Clarity is an ultra-sensitive microplate luminometer that combines sensitivity, ergonomic design, and data analysis. Applications include reporter gene assays (luciferase, dual luciferase), ATP assays (cell proliferation, cytotoxicity), DNA assays, reactive oxygen species assays, and chemiluminescence assays. Fast photon counting and high-quality optics ensure the best sensitivity available. The use of disposable tips for injection, a contamination-free stainless steel work surface, and a dead volume below 500 μl are just a few of the features.

Cayman Chemical
For more information
800-364-9897
www.caymanchem.com
www.scienceproductlink.org

Antioxidant Assay

This Antioxidant Assay is designed to measure the overall antioxidant capacity within a given sample. The assay relies on the ability of antioxidants in the sample to inhibit the oxidation of ABTS in comparison to Trolox, a water-soluble tocopherol analog. By quantifying the cumulative effect of all antioxidants present, more relevant biological information is acquired compared with the measurement of individual components alone. The 96-well plate format can be used for the rapid measurement of antioxidant capacity in a variety of sample types, including plasma, serum, urine, saliva, and cell lysates.

Genotech
For more information
314-991-6034
www.genotech.com
www.scienceproductlink.org

Genomic DNA

OmniPrep is for the extraction of high-quality genomic DNA from any species or tissue. The unique feature of OmniPrep is that the genomic DNA hydrates in minutes, not hours. OmniPrep generates pure genomic DNA through a two-step protocol that removes proteins and other impurities. The procedure takes as little as 30 min. The DNA extracted is on average 100 kb in size and has an A260/280 ratio between 1.8–2.0. The genomic DNA can be easily digested by many restriction endonucleases.

Techne
For more information
+44 (0) 1223 832401
www.techne.com
www.scienceproductlink.org

Real-Time Nucleic Acid Detection System

Quantica is a real-time nucleic acid detection system that offers open chemistry versatility and multiplex capability in which up to four dyes can be detected in any one well. Quantica makes use of a cold solid-state white light source that has an impressive range of excitation wavelengths that does not limit the user to the choice of methods possible, compared with those offered by single wavelength systems. It features a photon-counting photo multiplier tube detection system that has a wide detection range suitable for use with fluorochromes currently used in real-time detection techniques. It also features easily interchangeable filter sets and two methods of controlling light levels. Its optical heated lid can be set at different temperatures for program stages requiring special incubations.

Nexcelom Bioscience
For more information
978-397-1125
www.nexcelom.com
www.scienceproductlink.org

Micro-Ruled Coverslip

The Cellattice micro-ruled coverslip consists of a cell culture surface with microscopic identification and measurement markers. It is manufactured with high optical-quality plastic suitable for phase contrast and other standard cell-based assays measurement techniques. The coverslip thickness is 0.13 mm to 0.16 mm. The large micro-ruled area is 10 mm by 10 mm. Cells cultured on the Cellattice coverslip are identifiable within 25 μm. The built-in markers on the coverslip allow direct measurement of cell growth and movement without image acquisition. The same cell or cell cluster can be measured throughout the course of the experiment, even after each incubation period. The Cellattice coverslip can be used to monitor morphology, cell movement, and differentiation at the individual cell level. Cell proliferation is measured directly with multiple readings.

Promega
For more information
608-277-2545
www.promega.com
www.scienceproductlink.org

Reporter Vectors

The Promega Rapid Response Reporter Vectors are designed to allow scientists to measure what they might have missed in their cell-based assays. The reporter vectors reduce the risk of secondary effects and allow measurements of smaller and more transient changes in transcription compared with their predecessors. The reporter vectors contain degradation sequences that yield a dramatic improvement in the temporal coupling between transcription and reporter signal. The relative magnitude of reporter response is also greater. The combined effect produces greater changes in less time, enabling researchers to reduce assay times by up to 75%. They are available in firefly or *Renilla* luciferase configurations, and are compatible with a wide range of Promega luminescent detection reagents. Primary uses include quantitative cellular analysis and functional genomics applications.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and government organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier by visiting www.scienceproductlink.org on the Web, where you can request that the information be sent to you by e-mail, fax, mail, or telephone.